UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Metóda binárnej klasifikácie založenej na finančných mierach rizika

DIPLOMOVÁ PRÁCA

2020 Bc. Jakub Hrbáň

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

Metóda binárnej klasifikácie založenej na finančných mierach rizika

DIPLOMOVÁ PRÁCA

Študijný program:	Ekonomicko-finančná matematika a modelovanie
Študijný odbor:	1114 Aplikovaná matematika
Školiace pracovisko:	Katedra aplikovanej matematiky a štatistiky
Vedúci práce:	doc. RNDr. Beáta Stehlíková, PhD.

Bratislava 2020 **Bc. Jakub Hrbáň**

Univerzita Komenského v Bratislave
Fakulta matematiky, fyziky a informatiky

ZADANIE ZÁVEREČNEJ PRÁCE

Meno a priezvisko študenta: Bc. Jakub Hrbáň
Študijný program: ekonomicko-finančná matematika a modelovanie (Jednoodborové štúdium, magisterský II. st., denná forma)
Študijný odbor: aplikovaná matematika
Typ záverečnej práce: diplomová
Jazyk záverečnej práce: slovenský
Sekundárny jazyk: anglický

Názov: Metóda binárnej klasifikácie založenej na finančných mierach rizika
Binary classification method based on financial risk measures

Anotácia: V diplomovej práci sa vysvetľuje metóda klasifikácie založená na vzťahu klasifikácie a finančných mier rizika, ktorá bola navrhnutá v článku od J.Y.Gotoha, A.Takeda a R.Yamamota, s názvom Interaction between financial risk measures and machine learning methods. Táto metóda je autorom práce implementovaná a aplikovaná na reálne dáta. Medzi dátami sú známe testovacie sady dát používané pri testovaní klasifikačných algoritmov, ako aj nové vlastné dáta, ktoré sa týkajú zvolenej zaujímavej a aktuálnej problematiky.

Vedúci: doc. RNDr. Mgr. Beáta Stehlíková, PhD.
Katedra: FMFI.KAMŠ - Katedra aplikovanej matematiky a štatistiky
Vedúci katedry: prof. RNDr. Marek Fila, DrSc.

Dátum zadania: 07.01.2019

Dátum schválenia: 08.01.2019

prof. RNDr. Daniel Ševčovič, DrSc.
garant študijného programu

.....
študent

.....
vedúci práce

Poďakovanie Rád by som sa touto cestou poďakoval svojej školiteľke doc. RNDr. Beáte Stehlíkovej, PhD. za jej pomoc a odborné rady, ktoré som od nej po celý čas dostával. Ďakujem svojim rodičom, ktorí mi opäť raz vytvorili ideálne podmienky na kľudné a bezproblémové písanie práce. Osobité ďakujem patrí spolužiakovi Bc. Máriovi Igazovi, ktorý mi poskytol cenné konzultácie potrebné k praktickej časti práce. V neposlednom rade by som rád poďakoval svojím spolužiakom, za krásne prežitý čas počas celého spoločného štúdia na Matfyze.

Abstrakt v štátnom jazyku

HRBÁŇ, Jakub: Metóda binárnej klasifikácie založenej na finančných mierach rizika [Diplomová práca], Univerzita Komenského v Bratislave, Fakulta matematiky, fyziky a informatiky, Katedra aplikovanej matematiky a štatistiky; školiteľ: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2020, 83s.

V diplomovej práci sme sa zamerali na vytvorenie modelov určených na binárnu klasifikáciu dát. Modely sú kombináciou klasifikačnej metódy strojového učenia zvanej *Support vector machine* (SVM) a modelov na meranie finančného rizika *Conditional value-at-risk* (CVaR) a *Mean-absolute-semideviation* (MASD). Postupnými odvodeniami týchto kombinovaných modelov sa dostávame až k úlohe lineárneho programovania, ktorú riešime v softvéri R. V prvej časti práce sme sa zamerali na to, ako metódy CVaR a MASD obstáli v porovnaní s bežne používanými klasifikačnými metódami. Toto porovnanie sme spravili pre dva dátové súbory, v ktorých obe metódy dopadli porovnateľne s bežnými metódami, pričom v niektorých metrikách ich dokonca predčili. V druhej časti sme sa na odvodené modely pozreli ako na nástroj dátového pred-processingu. Použitie l-1 normy v úlohe lineárneho programovania totiž priradí mnohým váham v modeli nulovú hodnotu, čo môžeme vnímať ako výber vstupných premenných do iných modelov. Po otestovaní tejto aplikácie modelov, to znamená výberu len istej podmnožiny vstupných premenných, sme v jednej z dvoch situácií badali lepšie výsledky pri predikciách metódou logistickej regresie, voči prípadu, keď táto metóda počítala so všetkými vstupnými premennými naraz.

Kľúčové slová: Metóda oporného bodu, CVaR, strojové učenie, klasifikácia, lineárne programovanie

Abstract

HRBÁŇ, Jakub: Binary classification method based on financial risk measures [Master Thesis], Comenius University in Bratislava, Faculty of Mathematics, Physics and Informatics, Department of Applied Mathematics and Statistics; Supervisor: doc. RNDr. Beáta Stehlíková, PhD., Bratislava, 2020, 83s.

In this master thesis we focused on creating a binary data-classification models. These models are combination of classification method called Support vector machine (SVM) and the models to measure the financial risk called Conditional Value-at-risk (CVaR) and Mean-absolute-semideviation (MASD). By consecutive derivation of these combined models we get a linear programm that we solve in software R. In the first part of our thesis we focused on performance of our models CVaR and MASD in comparison to results of commonly used classification methods. This comparison we provided for two datasets for which both methods got similar results than commonly used methods and in some metrics they even outperformed them. In the second part we used the models as a tool in data pre-processing. By using l-1 norm in the linear program, it gives us sparse results for weights in our models. In this context the models can be understood as a feature selection models whose results will be consequently used as the input variables to other models. After testing this application, i.e. selecting only a subset of input variables as a result of our models, in 1 out of 2 cases we observed better results of logistic regression method in predictions in contrary when all input variables were used.

Keywords: Support vector machines, CVaR, machine learning, classification, linear programming

Obsah

Zoznam obrázkov

Zoznam tabuliek

Úvod

Žijeme v dobe s obrovskou mierou informovanosti. Dáta sú v súčastnosti zbierané z nespočetného množstva zdrojov, akými môžu byť naše smartfóny, počítače, meteorologické merače, rôzne senzory a veľké množstvo ďalších zariadení, s ktorými denne prichádzame do styku. Veľké firmy často získavajú z týchto zariadení obrovské množstvá dát, ktoré im ponúkajú rôznorodé informácie o správaní sa ich zákazníkov, predaji ich produktov, chode firmy, výnosoch či nákladoch. Na to, aby tieto dáta boli efektívne využité, existuje matematické odvetvie zaoberajúce sa analýzou dát, za pomoci klasických štatistických metód, ako aj metód strojového učenia. Vďaka týmto metódam sa tak dáta dajú efektívne využiť a prinášať firmám dodatočné zisky z ich činnosti. Jedným z príkladov by mohla byť analýza dát v oblasti energetiky a jej využitie na predikciu vývoja cien energií za účelom predaja a kúpy elektriny do budúcich období. Ďalším z príkladov môže byť hľadanie súvislostí medzi priateľmi na Facebooku a možným návrhom na nové priateľstvá či klasifikácia klientov v bankách podľa toho, ktorí budú platiči úveru a ktorí nie. Týmto spôsobom sa banka môže v budúcnosti vyvarovať finančným stratám spôsobených nesplateným úverom.

V našej práci sa budeme zaoberať poslednou z uvedených štatistických metód a tou je klasifikácia. Našim cieľom bude priblížiť klasifikačný model uvedený v článku [9], vysvetliť jednotlivé kroky tvorby daného modelu, naprogramovať ho vo vybranom softvéri a následne natrénovať a otestovať na dátach.

Klasifikačný model z vyššie spomenutého článku je kombináciou metódy strojového učenia zvanej metóda oporného bodu (z anglického *Support vector mechine* - SVM) a metódy merania finančného rizika zvanej *Conditional value-at-risk* so skratkou CVaR.

V úvodnej kapitole 1 sa zameriame na stručné, ale zrozumiteľné vysvetlenie, čo je to SVM metóda a ako a na čo sa používa v praxi. V druhej kapitole 2 si priblížime čo je to finančné riziko a metódy, akými sa dá merať, konkrétne metódu CVaR a MASD. V nasledovnej kapitole 3 opíšeme postupne vznik klasifikačnej metódy založenej na finančnej miere rizika navrhnutej v článku [9] a jej prevod na úlohu lineárneho programovania. V štvrtej kapitole 4 sme naprogramované úlohy CVaR a MASD otestovali na dátach o diabetese. Ich úspešnosť sme porovnávali s viacerými bežne využívanými metódami určenými na klasifikáciu. V piatej kapitole 5 sme sa pozreli na možné využitie modelov

CVaR a MASD na výber premenných a porovnali sme ich s ostatnými metódami, ktoré sa na túto oblasť dátovej analytiky používajú. V poslednej kapitole 6 sme na dátach z energetiky testovali klasifikačné modely CVaR a MASD ako aj ich možné využitie v oblasti výberu premenných. Všetky porovnávania, výpočty a grafické výstupy sme realizovali v softvéri R [37].

1 Klasifikačná metóda oporného bodu

Stále viac a viac sa v dnešnej dobe vo veľkých firmách zdôrazňuje agenda dátovej analytiky. Naprieč všetkými odvetviami vznikajú v spoločnostiach analytické tímy, ktorých kľúčovou úlohou je pochopiť enormné množstvo dát, ktorými firma disponuje a nejakým spôsobom ich zužitkovať. V oblasti dátovej analytiky je klasifikácia jedna z kľúčových tém, ktorá pomáha pri rozhodovacích procesoch či už manažérom podniku alebo aj samotným zamestnancom.

Ako bolo spomenuté v úvode, v tejto kapitole si bližšie uvedieme jednu z metód určenú na klasifikáciu dát a tou je klasifikačná metóda oporného bodu, z anglického *support vector classifier* (budeme používať skratku SVC), ktorá je podmnožinou väčšej triedy úloh zvaných *support vector machine* (SVM). Táto metóda bola vyvinutá komunitou zaoberajúcou sa počítačovou vedou v 90.tych rokoch minulého storočia [4]. Odvtedy metóda nabrala veľmi na popularite, najmä vďaka jej pomerne jednoduchému a intuitívnemu využitiu.

Zavedieme si teraz matematickú notáciu, ktorú budeme používať v celej práci, tak ako je uvedená v článku [9, s. 367], z ktorého vychádzame.

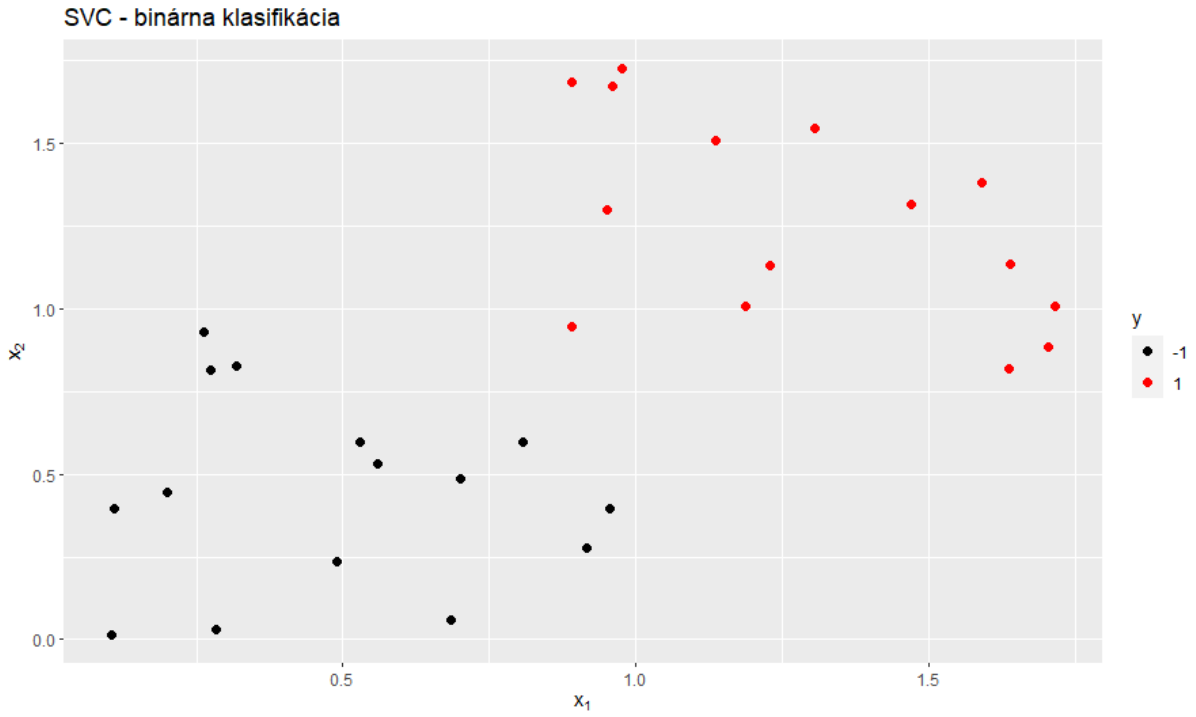
Nech $x \in \mathbb{R}^n$, potom $\|x\|$ značí ľubovoľnú normu v priestore $\mathbb{R}^n$. Vo všeobecnosti je l_p-norma definovaná nasledovne: $\|x\|_p := (\sum_{j=1}^{n} |x_j|^p)^{1/p}$ pre $p \in [1, \infty)$. Konkrétne l_1-norma (nazývaná aj poštárska alebo manhattanská) je definovaná ako $\|x\|_1 := \sum_{j=1}^{n} |x_j|$, l_2-norma (alebo aj euklidovská) ako $\|x\|_2 := (\sum_{j=1}^{n} x_j^2)^{1/2}$ a l_∞-norma (alebo aj nekonečnová norma) je definovaná ako $\|x\|_\infty := \max_{j=1,\dots,n} |x_j|$.

1.1 Lineárne separovateľný prípad

V tejto časti uvedieme a naformulujeme model SVC pre prípad klasifikácie do dvoch tried ako je uvedené v článkoch [9] a [11]. Úlohou SVC algoritmu je nájsť takú nadrovinu v n-rozmernom priestore, ktorá rozdeľuje súbor dátových bodov na dve skupiny. V prípade, že máme dve vstupné premenné, t.j. dáta pochádzajúce z priestoru $\mathbb{R}^2$, takouto nadrovinou bude priamka, ktorá dáta klasifikuje. V prípade, že dáta sú trojrozmerné, t.j. z priestoru $\mathbb{R}^3$, tak hľadanou nadrovinou bude 2D rovina v 3D priestore. Problém vizualizácie nastáva

od troch dimenzií vyššie. Naďalej však bude platiť, že dáta bude oddeľovať nadrovina s dimenziou o jedna menšou, ako má priestor, v ktorom sa nachádzajú.

Prejdením k matematickej notácii tak, ako je uvedená v [9], môžeme súbor našich dát označiť ako množinu m bodov $\{(x_1, y_1), ..., (x_m, y_m)\}$ v $\mathbb{R}^{n+1}$, kde $x_i \in \mathbb{R}^n$ sú namerané/získané veličiny $i-$teho pozorovania a $y_i \in \{-1, 1\}$ je zaradenie daného pozorovania do konkrétnej kategórie/triedy, kde $i = 1, ..., m$.



Obr. 1: Lineárne separovateľné dáta.

Na obrázku 1 možeme vidieť dáta rozmiestnené v priestore $\mathbb{R}^2$, ktoré sú na prvý pohľad lineárne separovateľné. Úlohou SVM algoritmu je nájsť takú separujúcu nadrovinu, ktorá maximalizuje svoju vzdialenosť od tzv. oporných bodov (z angl. support vectors), pričom táto nadrovina (definovaná vektorom váh $\mathbf{w} \in \mathbb{R}^2$ a afínnym posunom $\mathbf{b} \in \mathbb{R}$) by mala spĺňať nasledovné kritéria:

$$\begin{cases} y_i = +1 & \longrightarrow \mathbf{w}^T x_i > \mathbf{b} \\ y_i = -1 & \longrightarrow \mathbf{w}^T x_i < \mathbf{b} \end{cases} \tag{1}$$

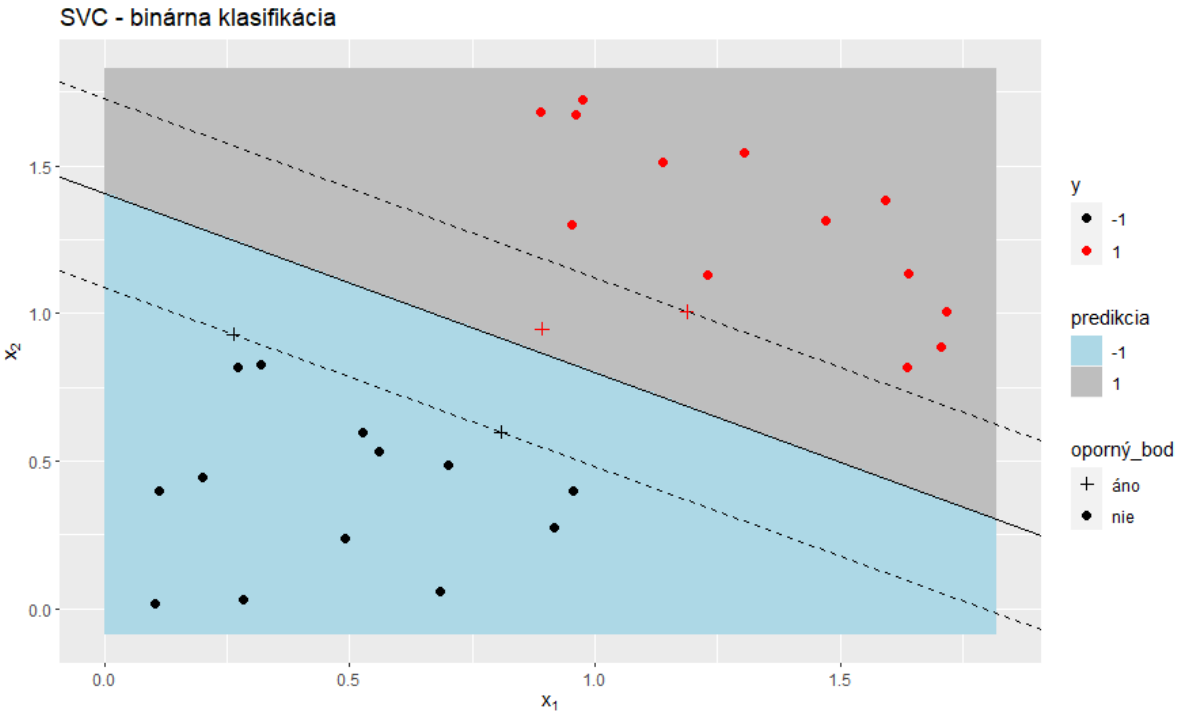
Je zrejmé, že tieto dve nerovnosti vieme zapísať ako $y_i(\mathbf{w}^T x_i - \mathbf{b}) > 0$, pre $i = 1, ..., m$. Optimalizačná úloha takéhoto problému je potom definovaná ako maximalizácia vzdiale-

nosti separujúcej nadroviny od bodov, respektíve vektorov, ktoré sú k nej v danej norme (v našom prípade euklidovskej) najbližšie. Formálny zápis takto naformulovanej úlohy uvádzame v (2).

$$\underset{w\in\mathbb{R}^n\setminus\{0\},b\in\mathbb{R}}{\text{maximize}} \min_{i=1,\dots,m} \frac{y_i(\mathbf{w}^T x_i - \mathbf{b})}{||\mathbf{w}||_2} \tag{2}$$

Vyššie uvedený zlomok $\frac{y_i(\mathbf{w}^T x_i - b)}{||\mathbf{w}||_2}$ predstavuje takzvané geometrické skóre, t.j. vzdialenosť bodu x_i od separujúcej nadroviny. V skutočnosti je len zopár bodov, ktoré určujú hranice separovanej nadroviny. Tie sa nazývajú oporné body.

Čo sa pri optimalizácii deje môžeme vidieť na obrázku (Obr. 2). Je očividné, že rovinou by mohla byť preložená ľubovolná priamka, ktorá oddelí obe kategórie, avšak metóda SVM našla práve takú separovanú nadrovinu, ktorá maximalizuje jej vzdialenosť od príslušných oporných bodov.



Obr. 2: Príklad separácie bodov optimálnou nadrovinou.

Vyššie spomenutú úlohu (2) vieme transformovať pomocou Charnes-Cooperovej transformácie [20] na úlohu konvexnej optimalizácie zvanú *Hard margin support vector classification* (HSVC) v nasledovnom tvare:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 \\ \text{s.t.} \quad & y_i(\mathbf{w}^T x_i - \mathbf{b}) \geq 1, \quad i = 1, ..., m \end{aligned} \tag{3}$$

Napriek tomu, že je úloha (2) zjednodušená na úlohu konvexného programovnia, stále dokáže riešiť len lineárne separovateľné dáta.

1.2 Lineárne neseparovateľný prípad

Teraz sme si uviedli príklad lineárne separovateľných dát v priestore $\mathbb{R}^2$, kde bolo ľahké určiť priamku, ktorá nám dáta pekne zakategorizuje. Čo však robiť v prípade, že dáta nie sú lineárne separovateľné? To znamená, že pre ľubovoľné dvojice $\mathbf{w}, \mathbf{b}$ existuje aspoň jedna dvojica (x_i, y_i) pre ktorú neplatí $y_i(\mathbf{w}^T x_i - \mathbf{b}) > 0$.

Ako sa uvádza v [9], jedným z riešení je často používané *C-support vector classification* (C-SVC) s nasledovným zápisom:

$$\begin{aligned} \min_{w,b,z} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 + \frac{C}{m}\sum_{i=1}^{m} z_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T x_i - \mathbf{b}) + z_i \geq 1 \\ & z_i \geq 0, \quad i = 1, ..., m, \end{aligned} \tag{4}$$

kde $C > 0$ je užívateľom definovaná premenná. Takto zadefinovaná úloha už má riešenie aj v prípade lineárne neseparovateľných dát.

V strojovom učení sa úloha tohto typu často zapisuje v tvare tzv. minimalizácie štrukturálneho rizika (structural risk minimization), pričom cieľom je minimalizácia dvoch objektov:

$$\min_{w,b} \quad \frac{1}{2}||\mathbf{w}||_2^2 + \frac{C}{m}\sum_{i=1}^{m} \max\Big\{1 - y_i(\mathbf{w}^T x_i - \mathbf{b}), 0\Big\}. \tag{5}$$

Prvý člen $||\mathbf{w}||_2^2$ predstavuje takzvaný regularizačný člen[1]. Druhý člen reprezentuje mieru nesprávnej klasifikácie a označuje sa ako empirické riziko. Zo zápisu môžeme vidieť, že

[1]Regularizačný člen sa často používa v modeloch z dôvodu predídeniu pretrénovania modelu na trénovacích dátach. Viac informácií sa nachádza v sekcii Príloha A v časti A.1.

do výpočtu priemeru nesprávnej klasifikácie vstupujú len tie dvojice pozorovaní (y_i, x_i), ktoré spĺňajú nerovnicu $y_i(\mathbf{w}^T x_i - \mathbf{b}) < 1$.

Pri takto navrhnutom modeli sa však vynárajú otázky ohľadom interpretovateľnosti niektorých jeho súčastí. Prvou z nich je interpretácia a určenie premennej C, ktorej hodnota sa neoptimalizuje, ale musí byť zadaná užívateľom predtým ako vstúpi do modelu. Druhou a zároveň podstatnejšou vecou je nejasnosť, prečo je model navrhnutý tak, že za nesprávne klasifikované dáta považuje aj tie, ktoré spĺňajú nerovnosti $0 < y_i(\mathbf{w}^T x_i - \mathbf{b}) < 1$ a nie len tie, ktoré spĺňajú $y_i(\mathbf{w}^T x_i - \mathbf{b}) < 0$.

S nápadom na takzvané *v-SVC* prišiel Schölkopf a kol. v roku 2000 [43]. Pridaním dodatočných premenných do modelu je zaručené, že model dokáže klasifikovať aj lineárne neseparovateľné dáta s istou dávkou nepresnosti. *v-SVC* (6) je nadstavbou úlohy (4), keďže z časti odstraňuje druhú z nejasností, ktoré vzišli ohľadom tohto modelu.

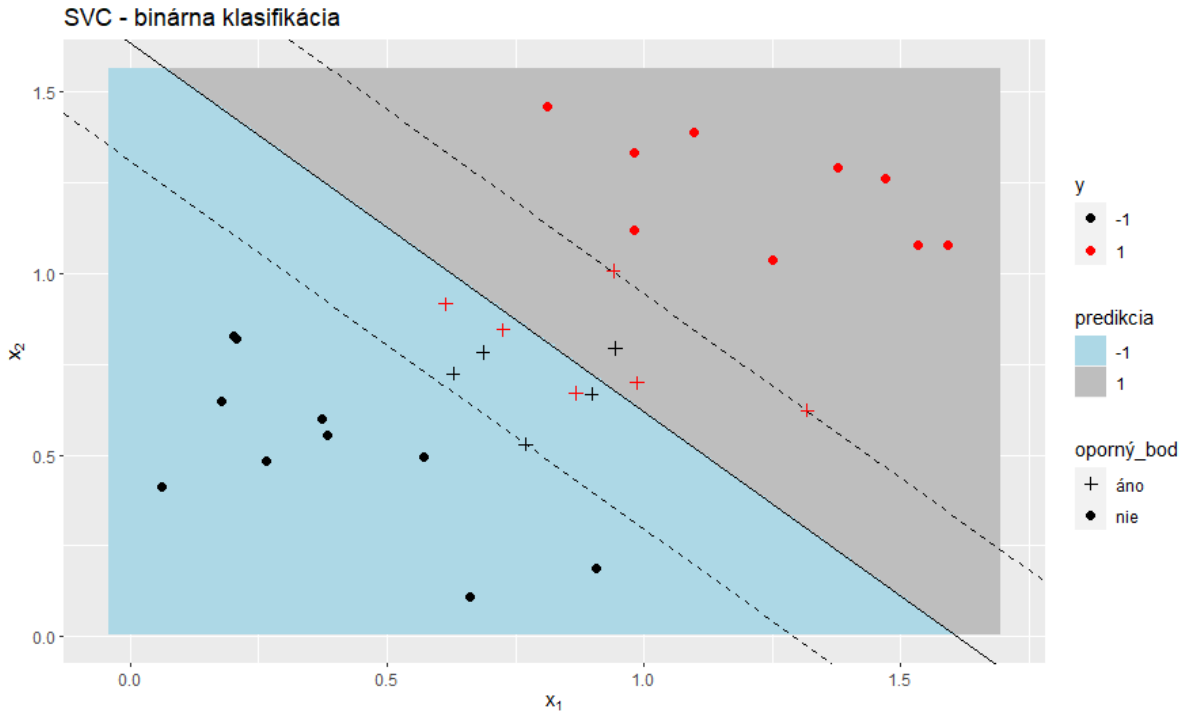
$$\begin{aligned} \min_{w,b,z,\rho} \quad & \frac{1}{2}\|\mathbf{w}\|_2^2 - \rho + \frac{1}{vm}\sum_{i=1}^{m} z_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^T x_i - \mathbf{b}) + z_i - \rho \geq 0 \\ & z_i \geq 0, \quad i = 1, ..., m, \end{aligned} \tag{6}$$

kde $v \in (0, 1]$ je voľne nastaviteľný parameter užívateľom a parameter ρ je pri dosiahnutí optimálneho riešenia nezáporný. *v-SVC* je opäť možné zápisať vo forme štrukturálneho rizika nasledovne.

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}||\mathbf{w}||_2^2 + \min_{\rho}\left\{ -\rho + \frac{1}{vm}\sum_{i=1}^{m} \max\left\{ -y_i(\mathbf{w}^T x_i - \mathbf{b}) + \rho, 0\right\}\right\} \tag{7}$$

Druhá minimalizačná časť úlohy sa dá interpretovať ako podmienená hodnota portfólia v riziku (z angl. *conditional value at risk*, označované skratkou *CVaR*), čo je prípadom koherentnej miery rizika vo financiách. O metóde CVaR ako aj o tom, čo znamená koherentná miera rizika a aké miery finančného rizika spĺňajú jej definíciu si viac povieme v kapitolách 2 a 3.

Uvádzame ešte jeden obrázok (Obr.3), na ktorom môžeme vidieť ako si SVM poradilo s lineárne neseparovateľnými dátami. Napriek tomu, že separujúca nadrovina nedokáže jednoznačne oddeliť dáta na kategórie $\{1, -1\}$, vidíme, že chyba pri prípadnej klasifikácii



Obr. 3: Lineárne neseparovateľné dáta.

nových bodov by bola v tomto ilustračnom príklade veľmi nízka.

2 Metódy merania finančného rizika

Téma finančného rizika je kľúčová naprieč všetkými oblasťami biznisu, počnúc bankovými inštitúciami až po malé výrobné firmy. Konkrétne sa zameriame na trhové riziko, ktoré je podmnožinou finančného rizika. Čo presne toto riziko predstavuje v oblasti rizikového manažmentu a spôsoby akými sa meria a vyhodnocuje si povieme v tejto kapitole.

Jednou z kľúčových úloh finančných inštitúcii po celom svete je identifikácia a minimalizácia rizík v rámci spoločnosti. Poznáme viaceré druhy rizík, medzi ktoré patrí aj práve spomenuté finančné riziko, konkrétne trhové riziko. Ide o typ rizika, kedy aktíva držané spoločnosťou náhle stratia na svojej trhovej hodnote. Najčastejšie to bývajú akcie a dlhopisy iných spoločností, avšak túto stratu môžu spôsobiť aj rôzne kurzové pohyby, v prípade, že inštitúcia drží prostriedky v inej mene. Takéto riziko by sa samozrejme dalo skúmať aj v nefinančných spoločnostiach, zaujímavé sú však práve finančné spoločnosti, ktoré vo väčšine prípadov držia a spravujú prostriedky svojich klientov a sú preto na nich vyvíjané isté požiadavky z radov dozorných inštitúcií akými sú napríklad americká centrálna banka (Federal reserve - FED), britská centrálna banka (Bank of England - BoI), európska centrálna banka (European Central Bank - ECB) alebo naša Národná banka slovenska (NBS).

2.1 Úvod do VaR

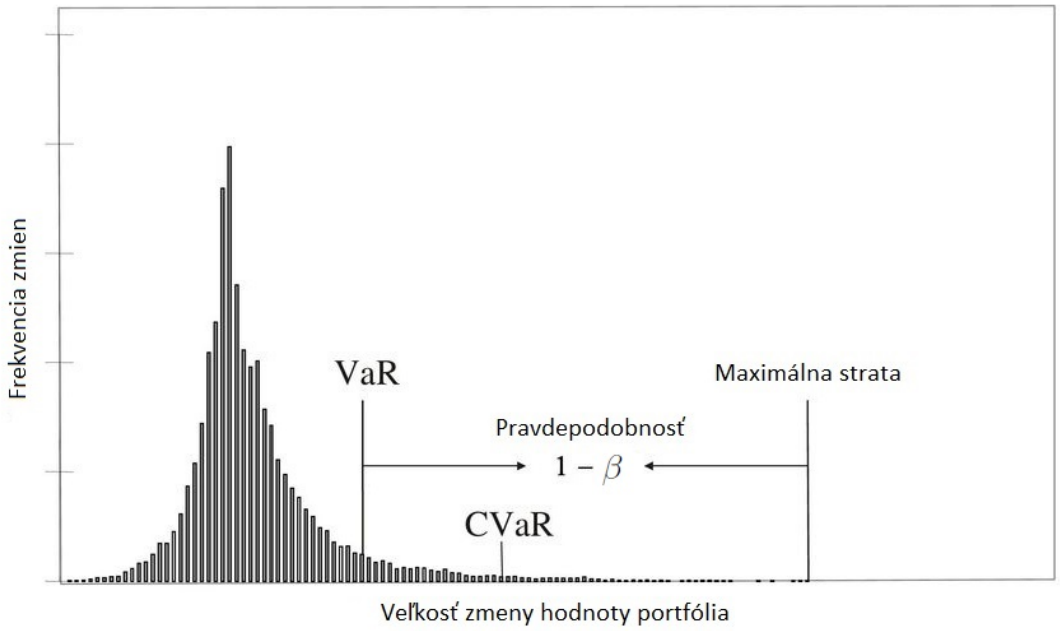
Metódy na výpočet trhového rizika boli vyvinuté už v 50.rokoch minulého storočia (viac sa dá dočítať v [16]), ale obmedzenosť výpočtovej techniky v tej dobe bránila pokročilejším modelom a tak väčšina z nich boli len teoretické. Až v 90.rokoch minulého storočia prišla spoločnosť RiskMetrics (spoločnosť pod bankou J.P.Morgan) s vlastným modelom na výpočet trhového rizika zvaným Value-at-Risk [16] (budeme používať skratku VaR). Tento model sa následne rozšíril do všetkých sfér finančného sektora a bol používaný či už veľkými bankami, poisťovňami a správcovskými spoločnosťami, ako aj mnohými ďalšími spoločnosťami obchodujúcimi na burze, ktoré si z dôvodu potreby mapovať riziko model VaR veľmi rýchlo osvojili. Výhodou je jeho jednoduchá implementácia na jednej strane skĺbená s dôležitou informáciou o možných rizikách na strane druhej.

Z hľadiska risk manažmentu je potrebné sledovať ako sa hýbe hodnota finančného

portfólia spoločnosti a následne modelovať nastatie možných strát portfólia. Model VaR je jedným z prostriedkov určených práve na takéto modelovanie. Funguje na základe pozorovania historického vývoja hodnoty finančného portfólia, presnejšie povedané pozorovania zmien jeho hodnoty. Pri konkrétnej metodike výpočtu VaR, metóda za určitej pravdepodobnosti povie, aká bude maximálna strata, ktorú v daný deň môže portfólio dosiahnuť. V praxi bežne používaným býva 99%-VaR. Ten hovorí, že s 99% pravdepodobnosťou strata hodnoty portfólia nepresiahne metódou vypočítanú hodnotu. Inak povedané, z pravdepodobnostného hľadiska sa jedná o 99% percentil z histogramu rozdelenia strát portfólia (v diskrétnom prípade). Hodnotu VaR, označovanú ako α, vieme formálne zapísať nasledovne:

$$\alpha_\beta[\mathcal{L}] := \min_\alpha\{\mathbb{P}\{\mathcal{L} < \alpha\} \geq \beta\}, \tag{8}$$

kde $\mathcal{L}$ je náhodná premenná z pravdepodobnostného priestoru $\mathbb{P}$ náhodných premenných predstavujúcich straty portfólia, hodnota $\beta \in (0, 1)$ nám určuje, pri akej pravdepodobnosti VaR počítame (táto hodnota sa volí). Na obrázku (Obr.4) môžeme vidieť histogram strát simulovaného portfólia a pre neho vypočítanú hodnotu VaR.



Obr. 4: Ukážka VaR a CVaR zo simulovaného portfólia. Zdroj: upravené z [47]

2.2 CVaR

Napriek tomu, že VaR je široko využívaným nástrojom na meranie finančného rizika, bolo mnoho protichodných názorov, že takto merané riziko nedáva úplný obraz o možných stratách portfólia, navyše z matematického hľadiska nejde o koherentnú mieru rizika (viac v podkapitole 2.4), ktorú by modely určené na meranie rizika mali spĺňať. V roku 2000 prišli dvojica Rockafellar a Uryasev [39] s novou metódou na výpočet rizika strát portfólia zvanou Conditional Value-at-Risk (budeme používať skratku CVaR), niekedy označovanou aj ako Expected Shortfall. Metóda CVaR je veľmi podobná výpočtu VaR, lenže v tomto prípade sa počíta stredná hodnota strát, v prípade, že pri β-% pravdepodobnosti, straty portfólia prekročia istú hodnotu α. Takto definovaná metóda navyše spĺňa všetky predpoklady koherentnej miery rizika. Myšlienku za metódou CVaR môžeme sledovať na obrázku (Obr. 4). Formálny zápis úlohy je nasledovný:

$$\Phi_\beta[\mathcal{L}] := \min_\alpha \left\{ \alpha + \frac{1}{1-\beta}\mathbb{E}[\max\{\mathcal{L}-\alpha, 0\}] \right\}, \tag{9}$$

kde $\mathbb{E}$ značí strednú hodnotu strát portfólia $\mathcal{L}$ po prekročení hodnoty α. Parameter β sa opäť hýbe v rozmedzí $(0, 1)$. Jednou zo zaujímavých vlastností, ktoré môžeme v prípade CVaR-u pozorovať sú tvary funkcie Φ, ktoré dostáva pre rozličné hodnoty β. Zoberme si m realizácií strát L_i náhodnej premennej $\mathcal{L}$, $i = 1, ..., m$. Pre hodnotu parametra $\beta = 0$ bude mať funkcia CVaR tvar $\Phi_0[\mathcal{L}] = \mathbb{E}[\mathcal{L}] = \sum_{i=1}^{m} p_i L_i$, čo reprezentuje očakávanú stratu portfólia. Naopak pre hodnotu $\beta \approx 1$ z dôvodu celkovej minimalizácie výrazu (9) a zároveň veľkosti druhého členu budeme požadovať, aby hodnota α bola rovná hodnote maximálnej straty portfólia, čím sa dosiahne nulovosť druhého člena. Funkcia Φ v tomto prípade nadobudne tvar $\Phi_\beta[\mathcal{L}] = \max_i\{L_i\}$, čo nám indikuje maximálnu stratu portfólia.

Pokračujúc podľa článku [9], jedným z dôležitých príkladov využitia daného modelu je minimalizácia empirickej verzie miery rizika. Predpokladajme, že máme k dispozícii m historických pozorovaní výnosov ($\boldsymbol{R_i} \in \mathbb{R}^p$), $p-$tice spoločností z vopred vybraného portfólia, kde $i = 1, ..., m$. $\boldsymbol{R}^{m\times p}$ je potom matica p spoločností s m historickými výnosmi. Ďalej definujme stratu tohto portfólia ako $L_i(\pi) = -\boldsymbol{R_i}^T\pi$, kde symbol π predstavuje váhy jednotlivých spoločností v portfóliu. Vychádzajúc z (9) vieme úlohu zapísať ako minimalizáciu empirického CVaR-u nasledovne:

$$\min_{\alpha,\pi} \alpha + \frac{1}{1-\beta} \sum_{i=1}^{m} p_i \max\{L_i(\pi) - \alpha, 0\}, \tag{10}$$

kde $\beta \in [0, 1)$ je voľne nastaviteľný parameter a p_i je pravdepodobnosť nastatia danej straty pre i-te meranie. Platí, že $p_i > 0$ spĺňa $\sum_{i=1}^{m} p_i = 1$. Zvyčajne sa toto pravdepodobnostné delenie volí uniformne, to znamená $p_i = \frac{1}{m}$, $i = 1, ..., m$.

2.3 MASD

Ako sa uvádza v [9], ďalšou možnou mierou rizika spĺňajúcou vlastnosť koherentnosti je takzvaná *Mean-absolute-semideviation* (budeme značiť ako MASD) definovaná ako:

$$\begin{aligned} r[\mathcal{L}] &= \mathbb{E}[\mathcal{L}] + \lambda \mathbb{E}[\max\{\mathcal{L} - \mathbb{E}[\mathcal{L}], 0\}] \\ &= \sum_{i=1}^{m} p_i L_i + \lambda \sum_{i=1}^{m} p_i \max\left\{L_i - \sum_{j=1}^{m} p_j L_j, 0\right\}, \end{aligned} \tag{11}$$

pre všetky $\lambda \geq 0$. Metóda MASD je koherentnou mierou rizika len pre $\lambda \in [0, 1]$, avšak stále reprezentuje mieru rizika straty portfólia aj pre hodnoty $\lambda > 1$.

Napriek tomu, že táto metóda nemá veľmi intuitívnu interpretáciu v mapovaní rizika strát portfólia tak, ako mali napríklad metódy VaR alebo CVaR, môžeme ju stále chápať ako istú mieru, ktorá zachytáva riziko vychýlenia sa strát od ich očakávaných hodnôt v závislosti od hodnoty parametra λ.

2.4 Definícia koherentnej miery rizika

V tejto sekcii sa budeme opierať o poznatky od Artznera z [1], ktorý vo svojej publikácii predstavuje 4 axiómy pre rizikové miery. Tvrdí, že tieto axiómy by mali byť dodržané v prípade, ak sa má riziko efektívne regulovať a manažovať. Následne rizikové miery, ktoré spĺňajú dané axiómy z definície 2.1 nazýva koherentnými.

Definícia 2.1. *Riziková miera r sa nazýva koherentnou, ak spĺňa nasledovné 4 axiómy:*

- *Axióma T. Translačná invariancia. $r[\mathcal{L} + a] = r[\mathcal{L}] + a$, pre všetky $a \in \mathbb{R}$.*
- *Axióma S. Subaditivita. $r[\mathcal{L}_1 + \mathcal{L}_2] \leq r[\mathcal{L}_1] + r[\mathcal{L}_2]$, pre všetky $\mathcal{L}_1$ a $\mathcal{L}_2$.*

- *Axióma P. Pozitívna homogenita.* $r[a.\mathcal{L}] = a.r[\mathcal{L}]$*, pre každé* $a \geq 0$*.*
- *Axióma M. Monotónnosť. Ak* $\mathcal{L}_1 \leq \mathcal{L}_2 \implies r[\mathcal{L}_1] \leq r[\mathcal{L}_2]$*.*

V publikácii [1] sa ďalej popisujú definované axiómy. Axióma T nám hovorí o tom, že riziko finančného inštrumentu s náhodnou premennou výnosov (strát) $\mathcal{L}$ navýšené o istú hodnotu a sa navýši práve o túto čiastku. Vysvetlenie axiómy S môže byť interpretované nasledovne: diverzifikácia rizika medzi viaceré investície nevytvára žiadne dodatočné riziko. Práve naopak, takáto stratégia môže celkové riziko znížiť. Axióma P implikuje, že riziko investícií je úmerné jeho veľkosti. Ak teda znásobíme naše portfólio, znásobí sa aj riziko predstavujúce stratu v portfóliu. Posledná axióma M je opäť intuitívna. Ak sú straty prvého portfólia nižšie ako straty druhého, riziková miera je taktiež nižšia v prospech prvého portfólia.

Z axióm S a P navyše vyplýva ďalšia vlastnosť a tou je konvexita rizikovej miery [14]. Tá sa dá zapísať nasledovne:

- *Vlastnosť. Konvexita.* $r[\lambda\mathcal{L}_1 + (1-\lambda)\mathcal{L}_2] \leq \lambda.r[\mathcal{L}_1] + (1-\lambda).r[\mathcal{L}_2]$*, pre všetky* $\mathcal{L}_1$*,* $\mathcal{L}_2$ *a* $\lambda \in [0,1]$*.*

Ako už bolo spomenuté, metóda VaR nie je koherentnou rizikovou mierou. Jej matematické obmedzenia badať pri axióme subaditivity, tá je zároveň kľúčovou pre konvexnosť a tým pádom aj koherentnosť rizikovej miery. Príklad uvedený v [6] dokazuje, že hodnota VaR vychádzajúca z kombinácie dvoch portfólii je väčšia ako súčet rizík pre portfólia jednotlivo. Tým sa však vyvracia vlastnosť konvexity.

3 Predstavenie modelov CVaR a MASD

V prvých dvoch kapitolách sme sa venovali na pohľad dvom veľmi odlišným oblastiam. Zatiaľčo v kapitole 1 sme uviedli klasifikačný model SVM pre neseparovateľné dáta, v kapitole 2 sme sa zaoberali definíciami a spôsobmi výpočtov dvoch koherentných mier finančného rizika. V [9] sa však uvádza, že existuje spojitosť medzi týmito na prvý pohľad nesúvisiacimi oblasťami.

3.1 Navrhovaný model CVaR

Uvažujúc metódu SVM zapísanú vo forme minimalizácie štrukturálneho rizika z príkladu (7) a zápis minimalizácie finančného rizika vo forme CVaR-u (10), po vykonaní nasledovnej transformácie:

$$\begin{cases} \rho & \longrightarrow -\alpha \\ v & \longrightarrow 1-\beta \\ \frac{1}{m} & \longrightarrow p_i, \quad i = 1, ..., m, \end{cases}$$

dostávame formuláciu úlohy:

$$\underset{w,b}{\text{minimize}} \quad \frac{1}{2}||\mathbf{w}||_2^2 + \min_{\alpha} \left\{ -\boldsymbol{\alpha} + \frac{1}{1-\beta} \sum_{i=1}^{m} p_i \max\left\{ -y_i(\mathbf{w}^T x_i - \mathbf{b}) - \boldsymbol{\alpha}, 0 \right\} \right\}. \tag{12}$$

Takto naformulovaná úloha potom môže mať zmysluplnú interpretáciu ako minimalizácia štrukturálneho rizika, pričom riziko je vyjadrené metódou CVaR, kde $-y_i(\mathbf{w}^T x - \mathbf{b})$ predstavuje stratu a p_i je pravdepodobnosť jej nastatia, $i = 1, ..., m$. Takýto model však v prípade lineárne neseparovateľných dát nevylučuje nulovosť váh $\mathbf{w}$. Preto sa zavedie obmedzenie pre veľkosť normy pre príklad (12).

$$\begin{aligned} \min_{w,b,\alpha} \quad & \boldsymbol{\alpha} + \frac{1}{1-\beta} \sum_{i=1}^{m} p_i \max\left\{ -y_i(\mathbf{w}^T x_i - \mathbf{b}) - \boldsymbol{\alpha}, 0 \right\} \\ \text{s.t.} \quad & ||\mathbf{w}||_2 = 1 \end{aligned} \tag{13}$$

Ďalej sa spomína, že autori algoritmus testujú na dátach súvisiacich s kreditným rizikom spoločností. Jedným z predpokladov, ktorý v článku uvádzajú je nezápornosť vstup-

ných premenných voči schopnosti spoločností splácať úvery (teda vysvetľujúcej premennej). Táto vlastnosť premenných je zabezpečená v procese predprípravy dát.

Pri optimalizácii sa autori rozhodli použiť namiesto euklidovskej l-2 normy zo zápisu (13) poštársku l-1 normu (pre pripomenutie, l-1 norma je súčet absolútnych hodnôt vektora, v našom prípade váh $\mathbf{w}$), čo ako uvidíme, prináša isté výhody. Ako sami autori spomínajú, voľbu l-1 normy zakomponovali do modelu vychádzajúc z poznatkov z ich predchádzajúceho článku [10], kde sa použitie tejto normy ukázalo ako vhodná alternatíva k riešeniu nelineárneho problému. Berúc do úvahy nezápornosť váh a použitie l-1 normy, ktorá dáva v súčte hodnotu 1, vieme tieto vzťahy zapísať matematicky ako $e_p^T w = 1$, kde $w \geq 0$ a e_p je jednotkový vektor dĺžky p. Využitím týchto vzťahov dostávajú nasledovnú úlohu lineárneho programovania:

$$\begin{aligned} \min_{w,b,\alpha,z} \quad & \boldsymbol{\alpha} + \frac{1}{1-\beta} p^T \mathbf{z} \\ \text{s.t.} \quad & \mathbf{z_i} + y_i(\mathbf{w}^T x_i - \mathbf{b}) + \boldsymbol{\alpha} \geq 0, \quad i = 1, ..., m, \\ & \mathbf{z} \geq 0, \quad \mathbf{w} \geq 0, \quad e_p^T \mathbf{w} = 1. \end{aligned} \tag{14}$$

Ako sa spomína v [9] l-1 norma je odporúčaná pri riešení úloh strojového učenia, keďže dáva takzvané „riedke“ riešenie[2], čo znamená riešenie s veľa nulovými hodnotami $\mathbf{w}$. Táto vlastnosť uvedenej metódy sa javí ako veľmi užitočná, keďže môže poslúžiť na výber premenných a uľahčiť tým jeden krok pri tvorbe modelu, respektíve výber týchto premenných môže poslúžiť ako vstup do iných modelov. O tom si viac povieme v kapitole 4.

Uvedieme si zápis (14) aj v maticovom tvare, ktorý nám neskôr pomôže pri tvorbe samotného modelu v prostredí jazyka R.

$$\begin{aligned} \min_{w,b,\alpha,z} \quad & (1, \frac{1}{1-\beta} p^T, 0_p^T, 0)(\alpha, z^T, w^T, b)^T \\ \text{s.t.} \quad & \begin{pmatrix} e_{m\times 1}, & I_{m\times m}, & Y_{m\times m} X_{m\times p}, & -y_{m\times 1} \end{pmatrix} \begin{pmatrix} \alpha \\ z \\ w \\ b \end{pmatrix} \quad \geq \quad 0_{m+p+2} \\ & \mathbf{z} \geq 0, \quad \mathbf{w} \geq 0, \quad e_p^T \mathbf{w} = 1, \end{aligned} \tag{15}$$

kde 0_p a 0_{m+p+2} sú vektory samých núl, $e_{m\times 1}$ je vektor samých jednotiek, $I_{m\times m}$ je matica identity, $Y_{m\times m}$ je diagonálna matica s prvkami vektora $y_{m\times 1}$ na diagonále a $X_{m\times p}$ je matica m pozorovaní s p vstupnými premennými.

3.2 Navrhovaný model MASD

Ku koncu kapitoly 3 sme si okrem CVaR modelu uviedli aj model MASD, ktorý je taktiež koherentnou rizikovou mierou a podľa definície preto môže byť použitý na mapovanie rizík. Takto navrhnutý problém (11) autori v článku [9] opäť prevádzajú na úlohu lineárneho programovania. Jej znenie bude potom nasledovné:

$$\begin{aligned} \min_{w,b,z} \quad & -\sum_{i=1}^{m} p_i y_i(\mathbf{w}^T x_i - \mathbf{b}) + \lambda p^T \mathbf{z} \\ \text{s.t.} \quad & \mathbf{z_i} + y_i(\mathbf{w}^T x_i - \mathbf{b}) - \sum_{j=1}^{m} p_j y_j(\mathbf{w}^T x_j - \mathbf{b}) \geq 0, \quad i = 1, ..., m, \\ & \mathbf{z} \geq 0, \quad \mathbf{w} \geq 0, \quad e_p^T \mathbf{w} = 1. \end{aligned} \tag{16}$$

Tak ako v prípade CVaR-u aj pre MASD model uvedieme jeho maticový zápis.

$$\begin{aligned} \min_{w,b,z} \quad & (-p^T Y X, p^T y, \lambda p^T)(w^T, b, z^T)^T \\ \text{s.t.} \quad & \begin{pmatrix} (I - e_m p^T) Y X, & -y + (p^T y) e_m, & I \end{pmatrix} \begin{pmatrix} w \\ b \\ z \end{pmatrix} \quad \geq \quad 0_{m+p+1} \\ & \mathbf{z} \geq 0, \quad \mathbf{w} \geq 0, \quad e_p^T \mathbf{w} = 1, \end{aligned} \tag{17}$$

pričom značenie aj rozmery uvedených matíc a vektorov sú tie isté ako v príklade (15).

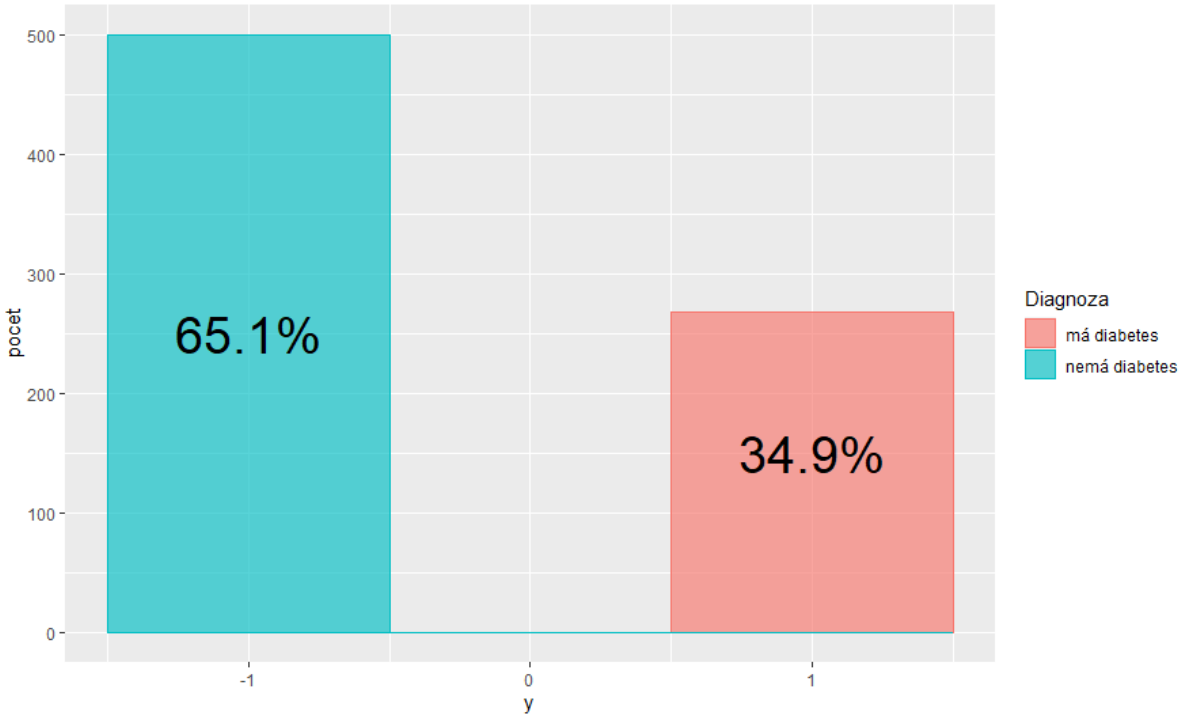
[2] l-1 norma je dobrým nástrojom na výber premenných, keďže pri optimalizácii pomerne vysoko penalizuje aj nízke hodnoty premenných (viac v sekcii A.2). Na výber premenných sa aktuálne používa aj takzvaná l-0 „norma“, ktorá udáva počet nenulových premenných. T.j. vo výraze $||x||_0 \leq r$ bude premenná $r \in Z_+$ ohraničovať maximálny počet nenulových premenných (viac v [34]).

4 Testovanie modelov na klasifikáciu

V nasledovných kapitolách prejdeme k praktickej časti diplomovej práce. Po stručnom oboznámení sa s metódou SVM a finančnými mierami CVaR a MASD v kapitolách 1 a 2 potrebných na odvodenie klasifikačného modelu navrhnutom v kapitole 3 teraz prejdeme na jeho vytvorenie v softvéri R (kódy je možne nájsť v B.1 a B.2). Model je v časti 4.6 testovaný a následne porovnávaný s inými známymi a bežne používanými štatistickými metódami a metódami strojového učenia.

4.1 Predstavenie dát

Súčasťou práce bolo nájsť vhodné dáta určené na klasifikáciu, na ktorých bolo možné modely natrénovať, otestovať a porovnať s ostatnými metódami. Rozhodli sme sa pracovať s dátovým súborom získaného z [25] pochádzajúceho z inštitútu NIDDK v Amerike, s dátami o ochorení diabetesu na populácii pôvodných obyvateľov Ameriky z kmeňa Pima. Špeciálne sa jedná o ženy v rozmedzí 21-81 rokov. Súbor obsahuje údaje o 768 ženách, pričom každý záznam obsahuje 8 vstupných premenných a jednu vysvetľujúcu premennú y, ktorá informuje o tom, či pozorovaná žena má ochorenie diabetesu.



Obr. 5: Rozdelenie vysvetľovanej premennej y.

Obrázok 5 ilustruje celkové počty pozorovaní v príslušných kategóriach ako aj percentuálne rozdelenie premennej y, ktorú sme pretransformovali na hodnotu $y \in \{-1, 1\}$, ktorá v tejto forme ďalej vstupuje do modelov (-1 = nemá diabetes, 1 = má diabetes).

Premenná y je modelovaná pomocou 8 zvyšných premenných. Nižšie si stručne uvedieme, čo ktorá premenná predstavuje:

- *pregnant* - počet, koľkokrát bola osoba tehotná,
- *glucose* - koncentrácia glukózy v krvi pri testovaní,
- *pressure* - krvný tlak v mmHg,
- *triceps* - hrúbka pokožky tricepsu,
- *insulin* - hladina inzulínu v krvi po testovaní,
- *mass* - hodnota BMI meraná ako $\frac{\text{váha}}{\text{výška}^2}$,
- *pedigree* - hodnota funkcie predispozície na chorobu,
- *age* - vek osoby v rozsahu 21-81 rokov.

4.2 Analýza a úprava dát

4.2.1 Úvodné štatistiky a čistenie dát

Pred samotnou implementáciou metód je vždy potrebné dáta upraviť a očistiť, aby algoritmy boli schopné dávať dobré výsledky. Keďže nami implementovaný algoritmus vie pracovať len s numerickými hodnotami vstupných premenných, potrebovali sme sa pozrieť na to, ako vyzerajú jednotlivé premenné. Pima dataset však obsahuje iba numerické dáta, preto nebolo potrebné robiť žiadne transformácie vstupných premenných. V ojedinelých prípadoch sa stáva, že vstupné premenné sú medzi sebou lineárne závislé, preto sme danú závislosť chceli overiť. Na túto operáciu bola použitá funkcia *findLinearCombos* z knižnice *caret* [28]. Dáta však pred nami už boli pravdepodobne starostlivo ošetrené a preto sme nenašli žiadne také premenné. Okrem toho sme zanalyzovali potenciálne NA hodnoty nachádzajúce sa v datasete. Tak ako v prípade linearity stĺpcov, aj teraz sme našli dokonale očistené dáta. Ak by sa v dátovom súbore predsa len nachádzali nejaké NA hodnoty, buď

by sme dané pozorovania odstránili z datasetu (v prípade malého počtu týchto hodnôt), alebo by sme pokračovali inou, vhodnejšou metódou, ktorá by sa snažila chýbajúce dáta nahradiť. Niektoré z klasifikačných modelov vedia pracovať aj s takými hodnotami v dátach, akými sú NA (napríklad rozhodovací strom), u nás by to však vyrobilo problém pri konštrukcii LP modelu, keďže do výpočtov vstupujú maticové číselné operácie. Základné štatistiky týkajúce sa datasetu je možné vidieť v Tabuľke 1.

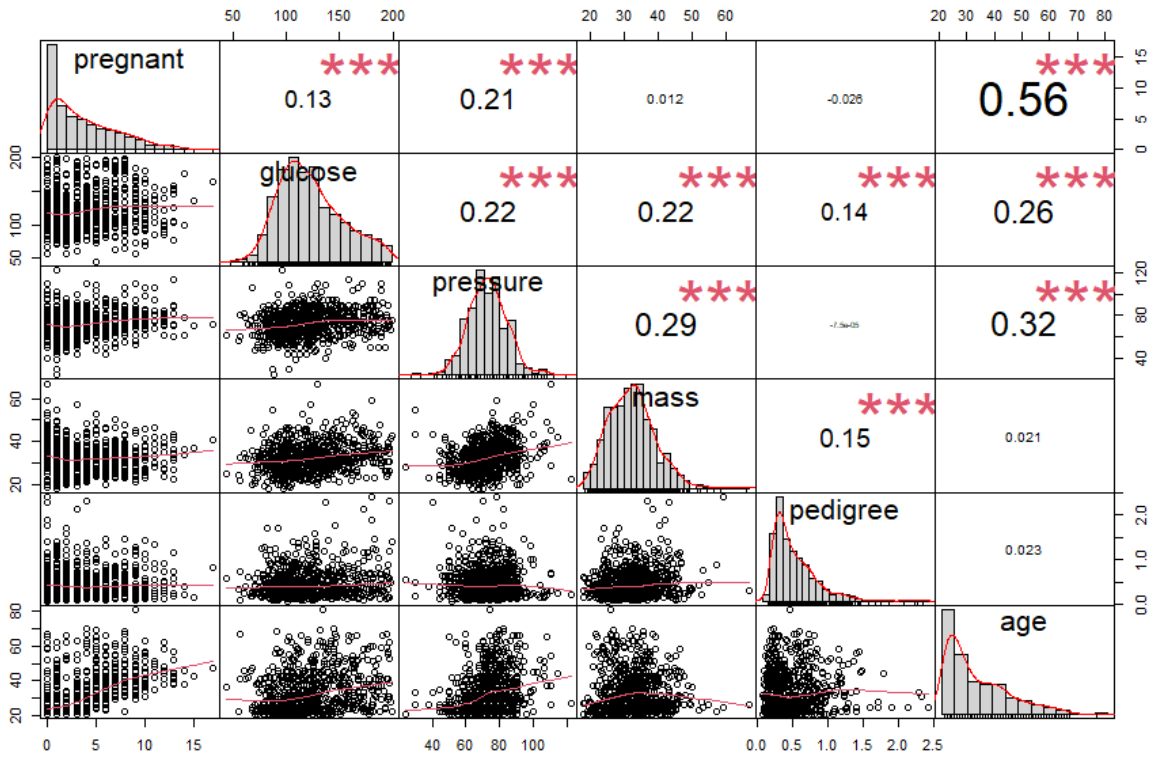
Premenné	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**NAs**	**Obs**	**NAsPerc**
pregnant	0.00	1.00	3.00	3.85	6.00	17.00	0.00	768	0.00
glucose	0.00	99.00	117.00	120.89	140.25	199.00	0.00	768	0.00
pressure	0.00	62.00	72.00	69.11	80.00	122.00	0.00	768	0.00
triceps	0.00	0.00	23.00	20.54	32.00	99.00	0.00	768	0.00
insulin	0.00	0.00	30.50	79.80	127.25	846.00	0.00	768	0.00
mass	0.00	27.30	32.00	31.99	36.60	67.10	0.00	768	0.00
pedigree	0.08	0.24	0.37	0.47	0.63	2.42	0.00	768	0.00
age	21.00	24.00	29.00	33.24	41.00	81.00	0.00	768	0.00

Tabuľka 1: Štatistiky vysvetľujúcich premenných datasetu.

Tabuľka 1 nám dáva základný náhľad na rozdelenie dát jednotlivých premenných. Ako sme už vyššie uviedli, dáta neobsahujú žiadne hodnoty NA, avšak čo môžeme sledovať, je výskyt nulových hodnôt v stĺpcoch, kde očakávame nenulové hodnoty a to konkrétne v stĺpcoch *glucose, pressure, triceps, insulin* a *mass*. Znepokojujúce sú navyše nulové hodnoty prvých kvartilov premenných *triceps* a *insulin*, čo znamená, že minimálne 25% meraní obsahuje nulové hodnoty. Pri bližšej analýze sme zistili, že až približne 300 pozorovaní z celkového počtu 768 je nulových v prípade premennej *insulin* a 200 pozorovaní z celkového počtu 768 v prípade premennej *triceps*. Práve z tohto dôvodu sme sa rozhodli tieto dve premenné z klasifikácie úplne vynechať. Čo sa týka nulových hodnôt zvyšných problémových premenných, tie boli podstatne nižšie (46 pozorovaní) a preto sme v ich prípade nulové pozorovania odstránili z datasetu. Dostávame tak finálny súbor obsahujúci 6 premenných s počtom pozorovaní 724.

Na obrázku 6 je možné pozorovať rozdelenie jednotlivých vstupných premenných ako aj korelácie medzi nimi. Vidíme, že dáta sú pomerne nízko korelované a preto môžeme predpokladať, že každá z premenných bude prispievať dôležitou informáciou do modelu.

Ak by naopak dáta boli medzi sebou vysoko korelované, či už kladne alebo záporne, každá z nich by prispievala len malým množstvom dodatočnej informácie.



Obr. 6: Histogramy premenných a ich vzájomné korelácie.

Po úvodnej analýze sme dáta rozdelili na trénovaciu a testovaciu zložku pomocou funkcie *createDataPartition* opäť z balíku *caret* [28]. Tá dáta rozdeľuje na základe kategórií vysvetľujúcej premennej. Túto funkciu je preto vhodné využiť v prípade, keď sa pracuje s nevyváženým datasetom, kde hrozí, že by napríklad funkcia *sample* nevybrala do testovacej vzorky ani jedno pozorovanie niektorej z kategórií. V literatúre sa uvádzajú rôzne pomery, v ktorých je dobré rozdeľovať dáta na trénovaciu a testovaciu, my sme zvolili pomer 80:20.

Dáta sme naškálovali na hodnoty v intervale $[0, 1]$ pomocou transformácie min-max, ako je uvedená v príklade (18). Na škálovanie testovacích dát sme taktiež použili min-max parametre trénovacej vzorky, keďže predpokladáme, že testovacie dáta pochádzajú z rovnakého rozdelenia ako trénovacie. Tým zároveň zabezpečíme väčšiu objektivitu a ešte

menšie zasahovanie do testovacej vzorky dát.

$$\text{naškálovaná premenná} = \frac{\text{premenná}_i - \min(\text{premenná})}{\max(\text{premenná}) - \min(\text{premenná})}, \quad i = 1, ..., 724 \qquad (18)$$

Tento proces sa vykoná pre všetky premenné. Hranice intervalu $[0, 1]$ sa dosiahnu v prípade, že premenná$_i$ je rovná minimálnej, respektíve maximálnej hodnote, ktoré premenná nadobúda.

4.2.2 Testovanie hypotéz vplyvu premenných

V úvode kapitoly 3 pri predstavovaní navrhovaného modelu sme uviedli, že v článku [9] predpokladajú nezápornosť premenných voči vysvetľovanej premennej. To znamená, že s narastajúcou hodnotou premennej by sa mala zvyšovať šanca, že pozorovanie padne do niektorej z dvoch kategórií. Interpretujúc to prostredníctvom našich dát o diabetese, tvrdenie by mohlo znieť, že s narastajúcim vekom pacientky (za predpokladu, že ostatné premenné sú konštantné a mení sa len vek) sa zvyšuje šanca, že pozorovaná žena má diabetes. V prípade, že by toto tvrdenie bolo pravdivé, máme zabezpečenú nezápornosť premennej voči vysvetľovanej premennej y a nemusíme ju nijakým spôsobom transformovať. Ak by sme však mali premennú, ktorá by s rastúcou hodnotou znižovala šancu, že daná pacientka má diabetes, to znamená, že na vysvetľujúcu premennú by vplývala negatívne, bolo by ju potrebné transformovať a to konkrétne prenásobením záporným znamienkom.

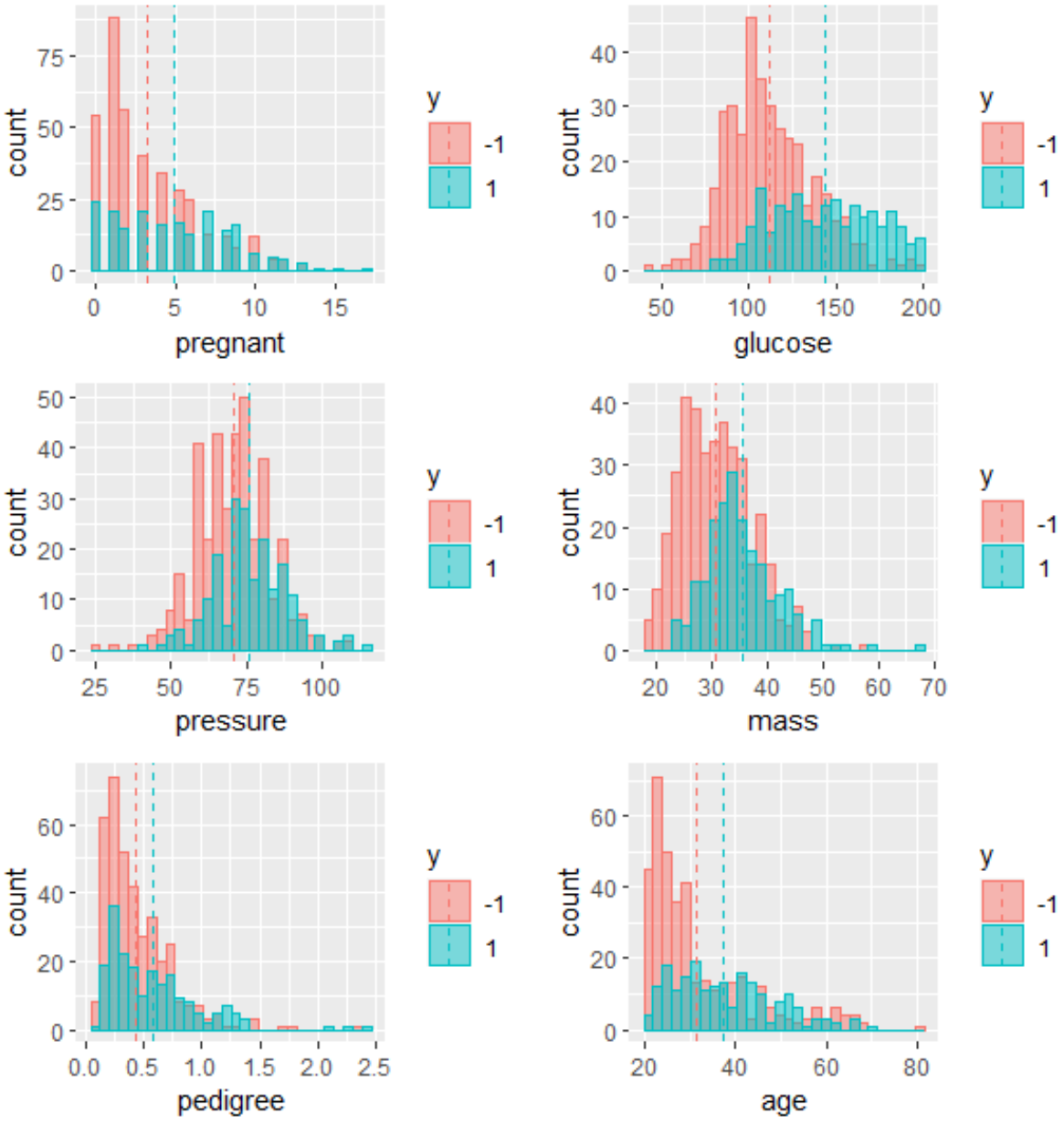
Spôsob akým sme testovali pozitívny / negatívny vplyv každej z premenných na vysvetľujúcu premennú, bolo štatistické porovnanie priemerov pozorovaní v závislosti od toho, do ktorej kategórie patria. Hypotéza H_0 tvrdí, že priemer μ_{neg} pre pozorovania, kde $y = -1$, je menší/rovný ako priemer μ_{poz} pre pozorovania s $y = 1$. Alternatívna hypotéza H_1 potom tvrdí opak. Formálny zápis takto uvedenej hypotézy bude nasledovný:

$$H_0 : \mu_{neg} \leq \mu_{poz} \quad vs. \quad H_1 : \mu_{neg} > \mu_{poz}$$

V prvom rade bolo potrebné zistiť normalitu rozdelení pozorovaní. To nám v R-ku zabezpečila funkcia *shapiro.test*, ktorá je jednou z funkcií softvéru R zo zabudovaného balíka *stats*. Na detekciu normality dát používa Shapirov-Wilkov test [42]. V závislosti od toho, či sa rozdelenia riadili normálnym rozdelením sa použil parametrický Studentov t-test [44], v opačnom prípadne neparametrický Wilcoxonov signed-rank test [52]. Použitie prvého

spomenutého testu zabezpečuje v softvéri R funkcia *t.test*, v prípade druhého *wilcox.test*. Obidve sú zo základného balíka *stats*.

Vo všetkých prípadoch testov na našich dátach bola hypotéza H_1 zamietnutá. To znamená, že každá zo vstupných premenných „rástla“ spolu s vysvetľovanou premennou a preto nebolo potrebné meniť znamienko žiadnej z nich. Na obrázku 7 môžeme pozorovať histogramy rozdelení jednotlivých premenných na základe zaradenia do tried $y \in \{-1, 1\}$. Prerušovanými čiarami sú na obrázkoch znázornené priemerné hodnoty rozdelenia danej premennej v závislosti od toho, do ktorej z dvoch kategórii patrí.



Obr. 7: Histogramy rozdelení vstupných premenných.

Napriem tomu, že sme postupovali ako bolo načrtnuté v článku [9] a snažili sa vyššie uvedeným štatistickým testovaním zistiť vplyv premenných na výslednú premennú, je dôležité pripomenúť, že takýto prístup nemusí byť vždy úplne vhodný. Vo výpočtoch a porovnávaniach priemerov totiž nie je zahrnutý vplyv ostatných premenných na vybranú premennú, ktorú práve testujeme. Na základe korelácií a našich testov sa nedá jednoznačne povedať či počet tehotenstiev má pozitívny vplyv na vznik ochorenia diabetu alebo je to v skutočnosti vplyv veku, keďže s rastúcim vekom logicky rastie aj možnosť porodiť väčší počet detí (v knihe [15] sa uvádza, že istý súvis medzi tehotenstvom a rizikom vzniku diabetu predsa len existuje). Testovanie preto môže byť v niektorých prípadoch ovplyvnené vonkajším faktorom, ktorý test neodhalí. Aj z toho dôvodu je dobré rozumieť dátam, s ktorými sa pracuje, pozrieť si výsledky testovania a kriticky zhodnotiť, či sú zmysluplné. V našom prípade sú všetky premenné, až na tehotenstvo, odrazom istej zdravotnej kondície pozorovanej osoby a preto výsledky testovania, to znamená ponechanie pozitívneho vplyvu premenných na vznik diabetesu, môžeme považovať za správne.

4.3 Porovnávacie kritériá

Ako hodnotiace a porovnávacie kritériá úspešnosti klasifikačných modelov sme vybrali tri metódy uvedené v [35]. Prvou z nich je metrika ACC (z anglického *accuracy*), ktorá meria podiel správne zaradených pozorovaní v testovacej vzorke voči celkovému počtu meraní. Nevýhodou tejto metódy však je, že nerozlišuje medzi chybou klasifikácie prvého a druhého druhu. To však môže byť kľúčové napríklad pri dátach o klinickom testovaní alebo schopnosti splácať úver, keďže v prvom prípade to môže mať zdravotné a v druhom finančné následky. Z toho dôvodu sme ako ďalšie dve metriky pozorovali hodnoty *Sensitivity* a *Specificity*, ktoré sa pozerajú na správnosť zaradenia testovaných pozorovaní do prvej, respektíve do druhej triedy.

Rozhodovaciu funkciu pre výpočet úspešnosti metriky ACC môžeme formálne zapísať nasledovne:

$$ACC(y, \hat{y}) = \frac{\sum_{i=1}^{m}(y_i == \hat{y}_i)}{m},$$

kde y je vektor hodnôt vysvetľovanej premennej testovacieho súboru, $\hat{y}$ je vektor hodnôt predikcií zvolenej metódy a m je počet pozorovaní zvoleného súboru dát. Uvedený zlo-

mok teda počíta do súčtu len tie pozorovania i, kde hodnota vysvetľovanej premennej a predikcie sa rovnajú.

V závislosti od zvolenej metódy sa $\hat{y}_i$ počíta rozličným spôsobom:

- **CVaR** - predikcia má tvar $\hat{y}_i = sign(\hat{w}^T x_i - \hat{b})$, kde x_i sú jednotlivé pozorovania a $\hat{w}$ a $\hat{b}$ sú parametre odhadnuté metódou CVaR,
- **MASD** - predikcia má tvar $\hat{y}_i = sign\Big((\hat{w}^T x_i - \hat{b}) - \sum_{i=1}^{m} \frac{(\hat{w}^T x_i - \hat{b})}{m}\Big)$, kde x_i sú jednotlivé pozorovania a $\hat{w}$ a $\hat{b}$ sú parametre odhadnuté metódou MASD. Druhý člen predstavuje priemernú hodnotu predikcií metódy CVaR.

		Merané hodnoty	
		-1	1
Predikcia	-1	True Negative	False Negative
	1	False Positive	True Positive

Tabuľka 2: Confusion matrix - matica sumarizujúca výsledky predikcie.

Pre výpočet ďalších dvoch metrík si najprv uvedieme pojem *confusion matrix* [35]. Jedná sa o maticu, ktorú dostávame po klasifikácii testovacej vzorky. Členy tejto matice nám hovoria o správnosti zaradenia daného pozorovania do príslušnej triedy. Ako taká matica vyzerá môžeme pozorovať v Tabuľke 2.

Zvyšné dve spomínané metriky majú potom nasledovné vzorce:

$$\text{Sensitivity} = \frac{\text{true negative}}{\text{true negative + false positive}},$$

$$\text{Specificity} = \frac{\text{true positive}}{\text{true positive + false negative}}.$$

Sensitivity hovorí o pomere správne zaradených meraní do prvej triedy danou metódou, voči všetkým meraniam, ktoré sú v skutočnosti z prvej triedy.

Specificity hovorí o pomere správne zaradených meraní do druhej triedy danou metódou, voči všetkým meraniam, ktoré sú v skutočnosti z druhej triedy.

4.4 Klasifikačné porovnávacie metódy

Na riešenie úloh lineárneho programovania, ktoré boli odvodené v maticovom tvare pre CVaR a MASD v príkladoch (15) a (17), sme použili v softvéri R balík s názvom *lpSolve* [2]. Tento balík rieši úlohu lineárneho programovania pomocou simplexovho algoritmu (viac o algoritme v [46]). Navrhnuté metódy tak boli porovnávané s 8 klasifikačnými metódami bežne využívanými v praxi. V Tabuľke 3 uvádzame balíky v softvéri R použité na ich výpočty.

Názov metódy	**Skratka**	**Použitý balík v softvéri R**
Logistická regresia [11] (*Logistic regression*)	LR	glmnet [8]
Metóda oporného bodu [11] (*Support vector machine*)	SVM	e1071 [36]
Lineárna diskriminačná analýza [11] (*Linear discriminant analysis*)	LDA	MASS [49]
Naivný Bayes [27] (*Naive Bayes*)	NB	klaR [51]
Neurónová sieť [19], [21] (*Neural network*)	NN	h2o [31]
Rozhodovací strom [11] (*Decision tree*)	DT	party [18]
Náhodný les [11] (*Random forest*)	RF	randomForest [33]
k-najbližších susedov [21] (*k-Nearest neighbours*)	kNN	class [49]

Tabuľka 3: Zoznam klasifikačných metód použitých na porovnanie s navrhnutým modelom.

Nižšie uvádzame stručný opis, ako metódy fungujú, pre prípad binárnej klasifikácie:

- *Logistická regresia* - namiesto priamej kategorizácie modeluje podmienenú pravdepodobnosť zaradenia meraní do príslušnej triedy. Keďže hodnoty pravdepodobností sú v intervale $(0, 1)$ je jednoduché určiť, po prekročení prahu (treshold), do ktorej

triedy dané meranie zaradíme. Pre ilustráciu uvažujme dáta o default-e spoločností. Potom ak $p(\text{default}) > 0.5$, kde p je pravdepodobnosť a 0.5 je nami určený treshold, označíme danú spoločnosť ako default-nú. Z dôvodu predídeniu pretrénovania logistickej regresie sme do modelu pridali regularizačný člen na koeficienty β s použitím l-2 normy. Logistická regresia s týmto rozšírením má pomenovanie *Ridge regression*. Mimo l-2 normy sa zvykne používať aj l-1 norma. V tom prípade dostávame takzvanú *Lasso regression*. Preferovaná je taktiež regularizácia v podobe lineárnej kombinácie týchto dvoch noriem. Takto navrhnutá metóda si vyslúžila označenie *Elastic net*. Základné informácie o vyššie uvedených nadstavbách logistickej regresie sú uvedené v [13].

- *Metóda oporného bodu* - túto metódu sme si podrobnejšie popísali v kapitole 1. Pre stručné zhrnutie, SVM metóda rozdeľuje priestor našich meraní pomocou separujúcej nadroviny, ktorej cieľom je oddeliť dáta na základe ich začlenenia do príslušnej triedy. V závislosti od toho, na ktorej strane separujúcej nadroviny sa bod/meranie nachádza, určí sa jeho trieda.

- *Lineárna diskriminačná analýza* - na rozdiel od logistickej regresie, ktorá modelovala podmienené rozdelenie vysvetľujúcej premennej y, LDA modeluje rozdelenie meraní pre každú z hodnôt vysvetľujúcej premennej a na základe priemerov a variancií vychádzajúcich z Bayesovskej vety vie určiť separujúcu nadrovinu, ktorá najlepšie kategorizuje dáta. Výhody oproti logistickej regresii spočívajú v stabilnejších odhadoch parametrov v prípade vysoko separovateľných dát a taktiež ak robíme viac triednu klasifikáciu.

- *Naivný Bayes* - názov naivný vychádza z predpokladu, že vstupné premenné sú navzájom nezávislé, čo pri reálnych dátach je skoro nereálne. Napriek tomu má táto metóda časté využitie v klasifikácii. Modelovanie je založené na výpočte pravdepodobností na základe Bayesovskej vety.

- *Neurónová sieť* - cieľom je modelovať fungovanie neurónov v našom mozgu. Vstupná informácia vo forme dát vojde do systému a na základe natrénovanej siete nám určí výsledok, v našom prípade výsledok klasifikácie. Opäť ako pri logistickej regresii aj tu sa aplikuje treshold na určenie finálnej predikcie. Použitá bola neurónová sieť

s jednou vnútornou vrstvou, ktorá bola tvorená 3 neurónmi. Za aktivačnú funkciu sme zvolili logistickú sigmoidu.

- *Rozhodovací strom* - ide o metódu, ktorá rozdeľuje priestor pozorovaní na viaceré „regióny“, pričom v každom určí jeho kategorické zaradenie. Určovanie týchto regiónov a nárast celého stromu funguje pomocou rekurzívneho binárneho delenia stromu, ktorého deliace podmienky sú nasledovného typu: ak je premenná $x > 25$, pozorovanie patrí do prvého regiónu, v opačnom prípade do druhého. Ukážku konkrétneho stromu si vyobrazíme v sekcii 4.6.

- *Náhodný les* - je metóda tvorená veľkým počtom rozhodovacích stromov, ktoré sú natrénované z dát, získaných bootstrap metódou (referencia). Finálna predikcia klasifikácie náhodným lesom sa určí väčšinovým hlasom výsledkov všetkých stromov v “lese„. Vytvorený bol les s použitím 500 rozhodovacích stromov.

- *k-najbližších susedov* - na základe vzdialeností bodov v priestore metóda určí k najbližších susedov referenčného bodu. Vzhľadom na triedy, do ktorej spadajú susedia daného bodu, sa pomocou väčšinového hlasovania určí, do ktorej triedy sa zaradí referenčný bod. V práci bola použitá kNN metóda s 9 susedmi, ktorá po krížovej validácii dosiahla najvyššiu hodnotu ACC.

4.5 Krížová validácia

Navrhnuté úlohy CVaR (15) a MASD (17) sú parametrizované a závisia od hodnoty β v prípade CVaR, respektíve od hodnoty λ v prípade metódy MASD, preto je potrebné zistiť, pre ktoré hodnoty týchto parametrov nám úlohy dávajú najlepšie výsledky. Na zistenie optimálnych hodnôt β a λ sme použili metódu k-zložkovej krížovej-validácie (z anglického *k-fold cross-validation*). Ide o metódu, kedy celú trénovaciu vzorku rozdelíme na $k-$podsúborov rovnakej veľkosti, pričom hodnotu k si vopred určíme. Proces následne prebieha tak, že pre každú hodnotu β, respektíve λ, sme vyčlenili $i-$ty podsúbor ako testovací a na zvyšných dátach sa model natrénoval, $i = 1, ..., k$. To sa zopakovalo pre každé i pričom v každom kroku sa vypočítala úspešnosť modelu pomocou metriky ACC. Tieto hodnoty úspešnosti sa následne spriemerovali a porovnali voči ostatným hodnotám

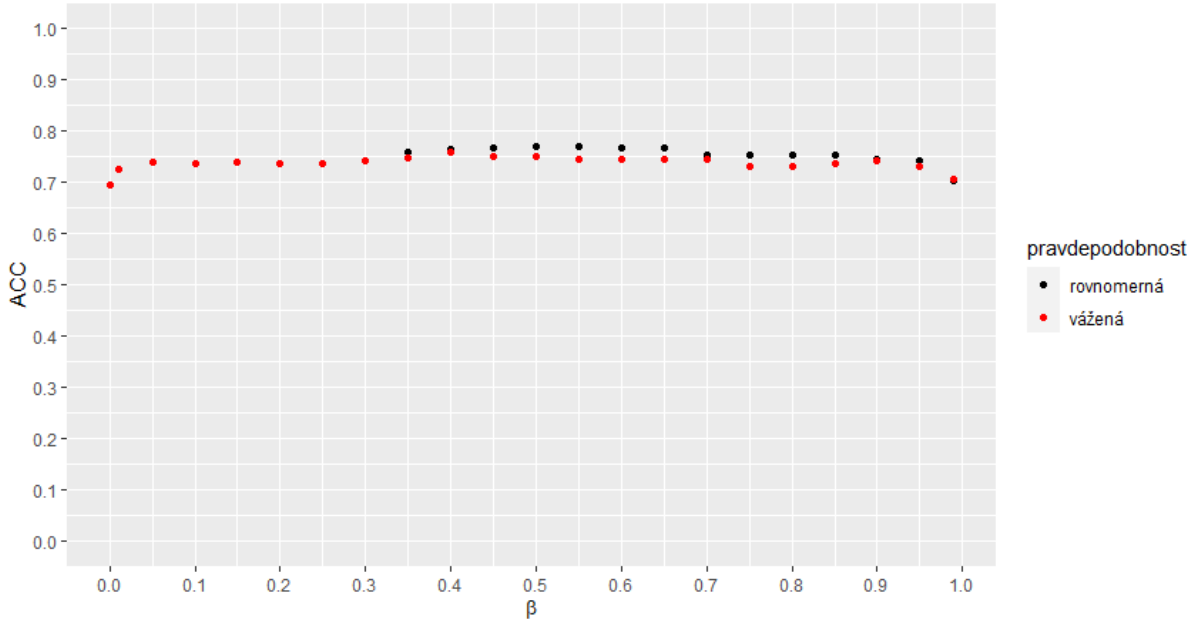
β, respektíve λ. Na záver, optimálne hodnoty β a λ boli tie, ktoré dosiahli najvyššiu priemernú hodnotu ACC.

Keďže hodnoty β v reči metódy CVaR predstavujú kvantily rozdelenia náhodnej premennej, boli vybrané nasledovné hodnoty parametra $\beta \in \{0, 0.01, 0.05, 0.1, ..., 0.95, 0.99\}$. Ako bolo už viackrát spomenuté, v metóde MASD namiesto parametru β vystupuje parameter λ, ktorý, ako bolo uvedené v podkapitole 2.3, nadobúda hodnoty $\lambda > 0$. Pre lepšiu interpretáciu sme túto hodnotu pretransformovali ako $\lambda = \frac{\tau}{1-\tau}$, kde $\tau \in [0, 1)$ nadobúda tie isté hodnoty ako β.

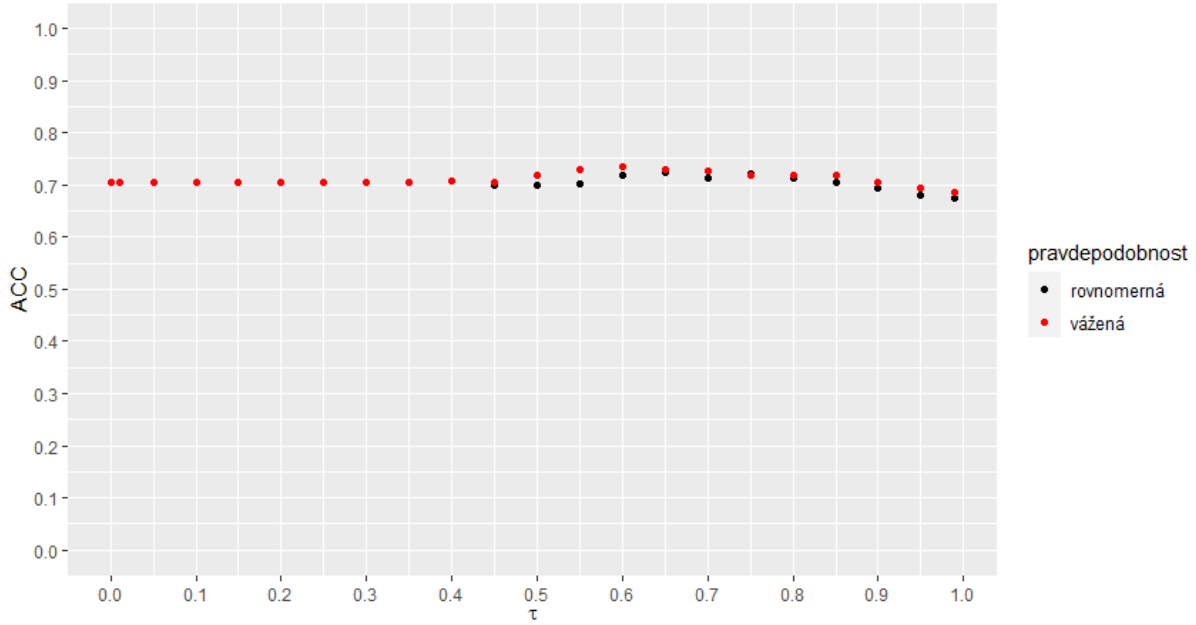
V úvode Kapitoly 3 sa vo formulácii pôvodnej úlohy (10) nachádza pravdepodobnosť p. Tak ako v článku [9] aj my sme uvažovali dve možnosti ako túto pravdepodobnosť určiť. Prvou z nich - *rovnomerná pravdepodobnosť*, je zvoliť $p = \frac{1}{m}$, kde m je počet pozorovaní trénovacej sady dát. Druhou z možností je zvoliť *váženú pravdepodobnosť*, kde sa hodnota p_i volí podľa zaradenia pozorovania do jednotlivých kategórií:

$$p_i = \begin{cases} \frac{1}{2m_+} \text{ pre } y_i = +1 \\ \frac{1}{2m_-} \text{ pre } y_i = -1, \end{cases}$$

kde m_+ značí počet tých pozorovaných osôb, ktorým bol diagnostikovaný diabetes a m_- tých, ktorým toto ochorenie nebolo diagnostikované.



Obr. 8: Výsledok krížovej validácie metódy CVaR pre rovnomernú a váženú pravdepodobnosť.



Obr. 9: Výsledok krížovej validácie metódy MASD pre rovnomernú a váženú pravdepodobnosť.

Na Obrázkoch 8 a 9 môžeme sledovať vývoj vypočítanej priemernej ACC v závislosti od hodnôt β a τ. Krížovú-validáciu sme robili pre dve rôzne pravdepodobnostné rozdelenia vstupujúce do modelu. Pre nízke hodnoty β a τ v prípade rovnomernej pravdepodobnosti pozorujeme, že nie sú vyobrazené žiadne hodnoty. Je to z dôvodu, že úloha lineárneho programovania hlási v týchto prípadoch neprípustnosť riešenia úlohy lineárneho programovania, zatiaľ čo v prípade váženej pravdepodobnosti simplexový algoritmus skonvergoval do optimálneho riešenia a bolo možné ACC vypočítať. Podobná situácia nastala aj v referenčnom článku [9].

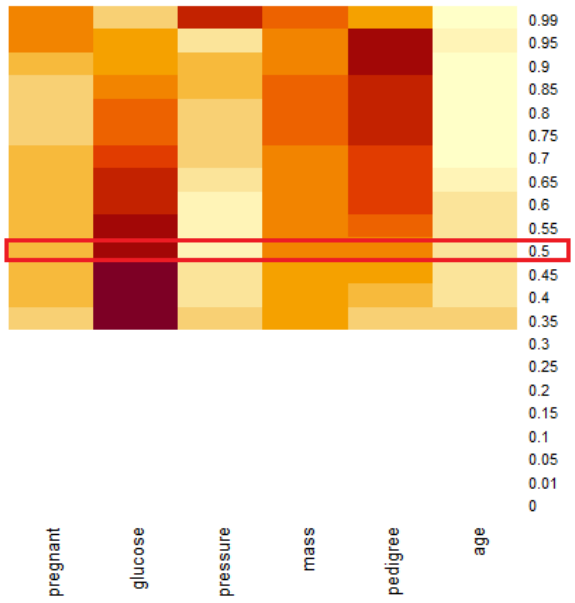
	Pravdepodobnostné rozdelenie	**Optimalizovaný parameter**	**Hodnota parametra**	**Accuracy**
CVaR	rovnomerné	β	0.50	0.7690
	vážené		0.40	0.7586
MASD	rovnomerné	τ	0.65	0.7241
	vážené		0.60	0.7345

Tabuľka 4: Výsledné hodnoty β a λ v závislosti od vybranej metódy a pravdepodobnosti.

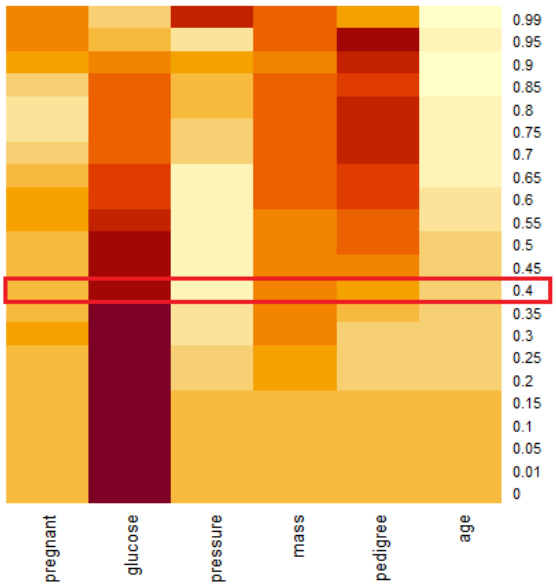
V Tabuľke 4 sa nachádzajú najvyššie dosiahnuté výsledky krížovej-validácie v metrike ACC na trénovacej vzorke pre všetky kombinácie uvedených metód a pravdepodobností, ako aj s uvedenými hodnotami β a τ, pri ktorých tieto ACC boli namerané.

Po identifikovaní optimálnych hodnôt β a τ sme opäť natrénovali model pre každú zo štyroch kombinácií na trénovacej sade, tentokrát na kompletnej vzorke, a otestovali na vyčlenenom testovacom dátovom súbore, ktorý zostal nedotknutý ešte z úvodného delenia. Výsledky sa zaznamenali a boli ďalej porovnávané s výsledkami predikcií ostatných metód uvedených v časti 4.4 na základe spomínaných metrík ACC, Sensitivity a Specificity.

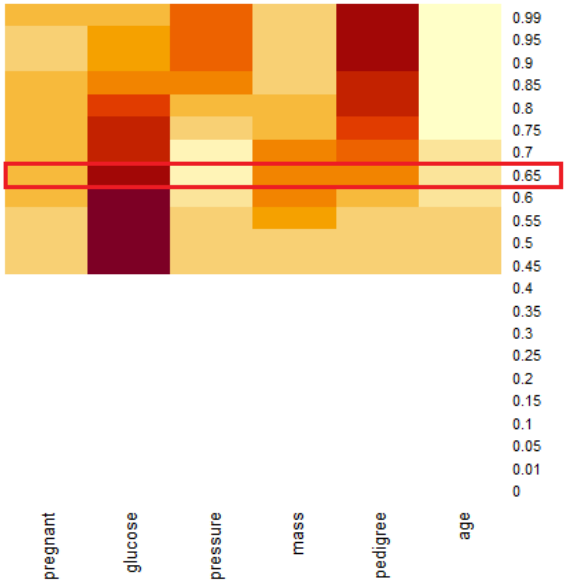
Ďalším výstupom z procesu krížovej validácie je vizualizáciu na Obrázku 10 v podobe heatmáp, ktoré vyobrazujú intenzitu výskytu vybraných premenných pre jednotlivé hodnoty β a τ metód CVaR a MASD. Platí, čím tmavší obdĺžnik, tým väčšia je váha premennej. Červenými obdĺžnikmi sú označené riadky s premennými a optimálnymi hodnotami β, respektíve τ, ktoré pri krížovej validácii dosiahli najvyššiu hodnotu ACC na validačnej dátovej vzorke. Z obrázkov môžeme pozorovať, že každá z konfigurácie metód CVaR a MASD zvolila za najdôležitejšie premenné *glucose* a *pedigree*, to znamená najväčšou mierou sa podieľajú na predikcii výskytu diabetesu. Ďalšími významnými premennými sa pri metóde CVaR javia stĺpce *pregnant* a *mass* a pri metóde MASD s vyššími hodnotami τ aj premenná *pressure*. Na Obrázkoch 10a a 10c vidíme, že pre nízke hodnoty β a τ nie sú uvedené žiadne intenzity premenných, čo je spôsobené tým, že pre tieto hodnoty obe metódy hlásili neprípustnosť. Naopak na Obrázkoch 10b a 10d pri nízkych hodnotách β a τ vidíme, že metódy skonvergovali, avšak vybrali len jednu premennú, konkrétne *glucose*. Táto vlastnosť nám napovedá, že pre nízke hodnoty parametrov β a τ majú modely tendenciu čo najväčšmi zjednodušiť trénovací proces a väčšinu premenných za pomoci l-1 normy posielajú do nuly. Deje sa tak z toho dôvodu, že člen optimalizačnej funkcie predstavujúci štrukturálne riziko menšou mierou prispieva k minimalizácii funkcie ako regularizačný člen. Z toho dôvodu metódy nechávajú len tú premennú, na základe ktorej vedia najlepšie zatriediť ženy do kategórie s alebo bez diabetesu. Že je to práve premenná *glucose* môžeme vyčítať aj z Obrázku 7, kde vidíme, že rozdiel priemerných hodnôt rozdelení pre jednotlivé kategórie má najväčšiu práve táto premenná a teda vie s najvyššou pravdepodobnosťou predpovedať kategórie voči ostatným premenným.



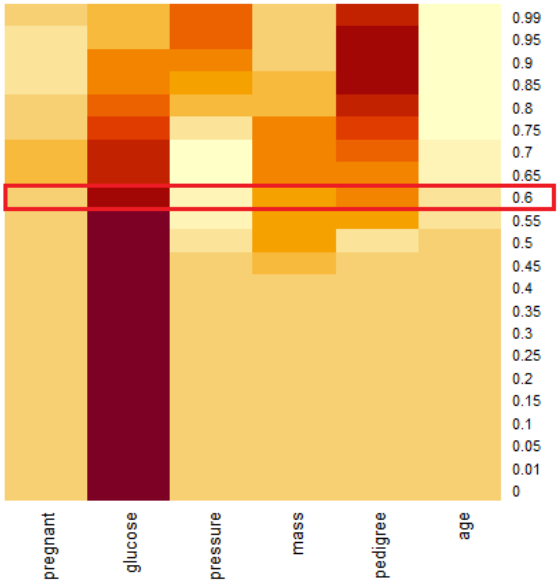
(a) Hodnoty β metódy CVaR s použitím rovnomernej pravdepodobnosti.



(b) Hodnoty β metódy CVaR s použitím váženej pravdepodobnosti.



(c) Hodnoty τ metódy MASD s použitím rovnomernej pravdepodobnosti.



(d) Hodnoty τ metódy MASD s použitím váženej pravdepodobnosti.

Obr. 10: Heatmapy znázorňujúce intenzitu premenných v modeloch CVaR a MASD pre rôzne hodnoty parametrov β a τ ako aj pravdepodobnostné rozdelenia.

4.6 Výsledky a porovnania

Výsledky našich 4 zvolených metód, ktoré sú počítané simplexovou metódou lineárneho programovania, sme sa snažili porovnávať s bežne používanými klasifikačnými metódami z Tabuľky 1. Ako sme už viackrát uviedli v tejto kapitole, metódy sú porovnávané na základe metrík ACC, Sensitivity a Specificity. Tie nám dokážu ukázať rôzne pohľady na spôsob klasifikácie porovnávaných metód.

Metódy	**Accuracy**	**Sensitivity**	**Specificity**
CVaR_r[3]	0.7361	0.8936	0.4400
CVaR_v	0.7361	0.8298	0.5600
MASD_r	0.7292	0.6915	0.8000
MASD_v	0.7569	0.7128	**0.8400**
LR	0.7500	**0.9574**	0.3600
SVM	0.7361	0.9149	0.4000
LDA	0.7500	0.9149	0.4400
NB	0.7153	0.8404	0.4800
NN	0.7639	0.8723	0.5600
DT	0.7292	0.7979	0.6000
RF	0.7500	0.8617	0.5400
kNN	**0.7708**	0.8936	0.5400

Tabuľka 5: Výsledné porovnania metód.

Z porovnania výsledkov Tabuľky 5 vidíme, že podľa metriky ACC vyšla najlepšie hodnotená metóda kNN. LR dosahuje síce priemerné výsledky ACC, avšak v metrike Sensitivity dosahuje hodnotu až 0.9574, čo znamená, že vo vyše 95% prípadov správne predikuje triedu -1, ktorú reprezentujú osoby, ktorým nebol diagnostikovaný diabetes. Vysoké skóre v metrike Sensitivity je však „vyvážené“ najnižšou hodnotou v metrike Specificity naprieč zvyšnými metódami. Metóda CVaR_r navrhnutá v článku [9] dosahuje priemerné výsledky spomedzi ostatných metód, čo sa týka metriky ACC a mierne nadpriemernú hodnotu Sensitivity. Sledujúc však metriku Specificity vidíme, že metóda CVaR dosahuje

[3]Indexy r a v pri metódach CVaR a MASD značia rovnomerné, respektíve vážené pravdepodobnostné rozdelenie použité v modeloch.

skôr podpriemerné výsledky. Metóda CVaR_v dosahuje rovnakú hodnotu metriky ACC, avšak opačne sa správa pri hodnotách metrík Sensitivity a Specificity. Veľkú zmenu však pozorujeme pri ďalších dvoch zvolených metódach MASD_r a MASD_v. Zatiaľčo hodnoty Sensitivity sú ďaleko najhoršie spomedzi všetkých ostatných metód, hodnoty Specificity naopak dosahujú najlepšie výsledky naprieč všetkými metódami. To môžeme vnímať pozitívne v prípade, ak by sme chceli maximalizovať práve túto hodnotu (respektíve minimalizovať chybu 1. druhu). V kontexte dát, s ktorými pracujeme, je minimalizácia práve tejto chyby žiadúca. Teda je lepšie minimalizovať počet tých, ktorí majú diabetes, ale test ich zaradí medzi zdravých ľudí a neliečia sa. Iný pohľad je taký, že pokiaľ viem s vysokou pravdepodobnosťou vyfiltrovať ľudí, ktorí nemajú diabetes, ostatným sa môže zabezpečiť náročnejšie testovanie (finančne aj časovo) a nemusia sa tak testovať všetci.

Spoločným problémom všetkých 4 navrhovaných metód je čas výpočtu. Samostatné metódy CVaR a MASD robia výpočet porovnateľne rýchlo ako napríklad LR alebo SVM. Problém však nastáva pri trénovaní, kedy musíme prehľadávať všetky riešenia pre hodnoty β, respektíve λ formou krížovej validácie. Tento proces nám vie v niektorých prípadoch výpočet predĺžiť až 20-násobne. Zároveň si však treba uvedomiť, že cieľom tejto diplomovej práce nie je testovať a optimalizovať výpočtovú náročnosť navrhovaných metód, preto túto informáciu treba brať len ako informatívnu, prípadne ako možnosť ďalšieho vývoja uvedenej triedy úloh.

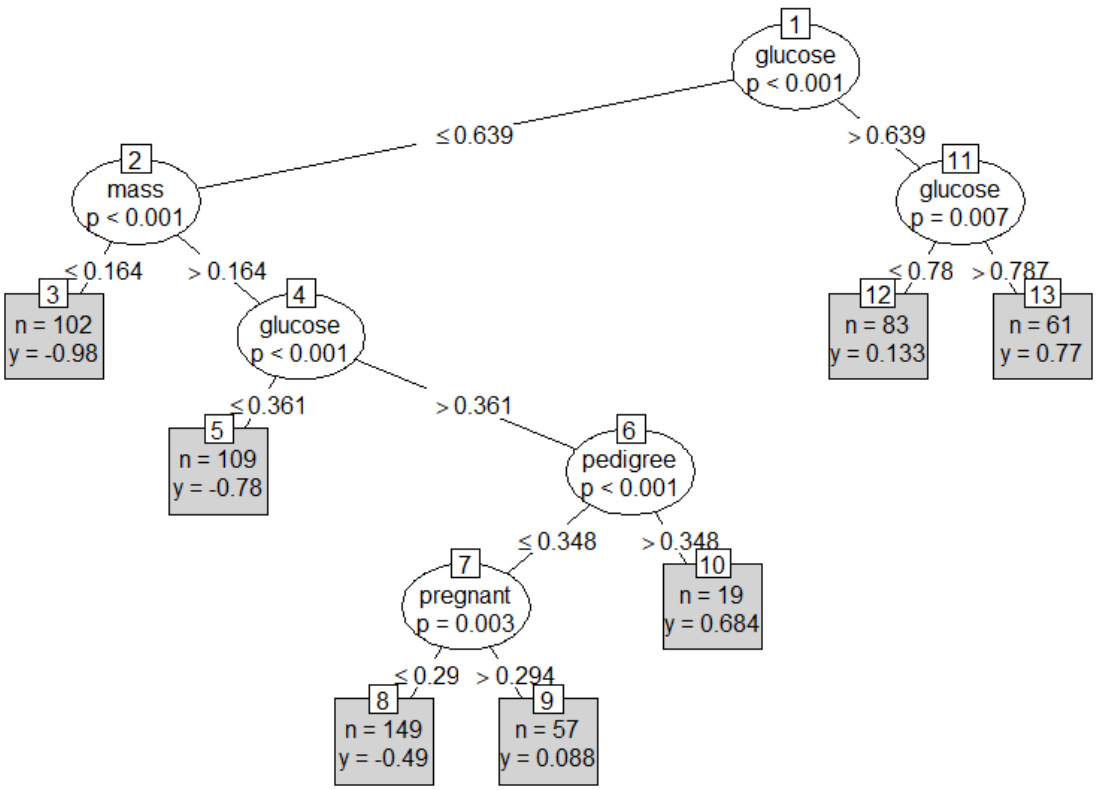
Keďže sme si ku koncu kapitoly 3 povedali, že model by mohol byť použitý aj na výber premenných, pozreli sme na to, ktoré premenné modely CVaR a MASD vybrali a porovnali sme ich s vybranými koeficientami premenných v logistickej regresii, s premennými SVM a LDA modelov.

Z Tabuľky 6 pozorujeme, že spomedzi metód CVaR a MASD len metóda CVaR_r vybrala všetky premenné z pôvodného datasetu. Výstupné hodnoty metód, sú hodnoty vektora váh w v pôvodne formulovanej úlohe. Zvyšné 3 metódy zvolili nulové váhy buď pre jednu alebo dve premenné. Sú to práve tie premenné, ktoré heatmapy na Obrázku 10 určili za menej významné pre všetky hodnoty β a τ. To je zapríčinené práve voľbou l-1 normy použitej v modeloch, ktorá penalizuje nízke hodnoty váh premenných. Pri logistickej regresii uvádzame koeficienty β, ktorých hodnota udáva tvar funkcie logistickej sigmoidy $\frac{1}{1+\exp^{-(\beta_0+x^T\beta)}}$, na základe ktorej sa model rozhoduje, kam zaradí dané meranie. Pre me-

	pregnant	**glucose**	**pressure**	**mass**	**pedigree**	**age**
CVaR_r	0.1257	0.3096	0.0036	0.2947	0.2315	0.0350
CVaR_v	0.1405	0.4468	-	0.2478	0.1096	0.0554
MASD_r	0.1452	0.3519	-	0.2798	0.2124	0.0107
MASD_v	0.1740	0.3999	-	0.2492	0.1769	-
LR	1.1879	3.0461	0.4553	1.8106	1.0502	0.7140
SVM	1.6026	3.8520	-0.0321	2.3955	1.4640	0.0680
LDA	1.5857	4.2613	-1.0022	2.2421	1.8548	0.8447

Tabuľka 6: Porovnanie koeficientov metód CVaR_r, CVaR_v, MASD_r, MASD_v, LR, SVM a LDA

tódy SVM a LDA koeficienty značia váhy w separujúcej nadroviny. Tak ako všetky metódy CVaR a MASD aj zvyšné 3 metódy prikladajú najväčšiu váhu na premenné *glucose* a *mass*. Zároveň môžeme sledovať, že premenné *pressure* a *age*, ktoré majú v LR, respektíve SVM a LDA, záporné alebo veľmi nízke váhy, naše metódy buď nezahrnú do výpočtov



Obr. 11: Vizualizácia rozhodovacieho algoritmu metódy decision tree.

vôbec alebo v prípade CVaR_r im priradia nízke váhy. Dovolíme si taktiež pripomenúť, že v pôvodnej analýze významnosti premenných metód CVaR a MASD, z Obrázku 10, sa po premennej *glucose* zdala byť druhou najvýznamnejšou premennou *pedigree*, avšak pri výbere optimálnej hodnoty β, respektíve τ, je to práve premenná *mass*, ktorá nadobúda druhú najvyššiu významnosť.

Na Obrázku 11 uvádzame schému rozhodovacieho algoritmu metódy *decision tree*, ktorý je veľmi pekne interpretovateľný. Zo štruktúry rozhodovacieho stromu vidíme, že metóda prikladá dôležitosť predovšetkým premenným *glucose* a *mass*, tak ako aj ostatné z metód v Tabuľke 6, a pri procese rozhodovania vynecháva premenné *pressure* a *age*. Ďalšie delenie by totiž model už nezlepšilo.

Zvyšné z metód určených na porovnávanie s modelmi CVaR a MASD nemajú jednoznačnú interpretáciu, čo sa týka či už vizualizácie alebo porovnávania váh modelov, preto pre ne uvádzame len hodnoty porovnávacích kritérií z Tabuľky 5.

5 Výber premenných

Jednou z vlastností, ktorá sa počas odvodzovania klasifikačných modelov (14) a (16) v článku [9] ukázala ako relevantná a prospešná je schopnosť modelov vybrať len určité premenné, pomocou ktorých sa ďalej počíta predikcia. Jedná sa o súčasť dátového preprocessingu (predprípravy dát), v ktorom sa pokúšame znížiť dimenziu počtu vstupných premenných, čím sa snažíme zabezpečiť vyššiu úspešnosť trénovaného modelu pri predikciách, nižšiu výpočtovú náročnosť (pri veľkom počte dát a komplikovanosti zvoleného modelu čas výpočtu zohráva kľúčovú rolu) a v neposlednom rade aj lepšiu interpretovateľnosť modelu. Motivácia použitia danej techniky býva v oblastiach, kde máme k dispozícii veľké množstvo vstupných premenných ako napríklad pri medicínskych dátach, spracovaní textových dokumentov, či rôznych fyzikálnych a chemických meraniach.

5.1 Prehľad metód

Sú tri okruhy metód (viac o metódach v [24],[32]), ktoré sa používajú na výber premenných (názvy uvádzame v angličtine, keďže sme k nim nenašli vhodný slovenský ekvivalent):

- *Filter methods* - ide o metódy, ktoré nezávisia od modelu, ktorý bude na dátach v ďalšej fáze použitý. Poväčšine ide o štatistické metódy, ktoré zisťujú dôležitosť premenných v dátovom sete či už vzhľadom na vysvetľovanú premennú alebo ostatné vysvetľujúce. Príkladmi takejto metódy môže byť sledovanie vzájomnej korelácie premenných, testy signifikantnosti alebo vzdialenosti bodov vrámci rovnakých klasifikačných tried.

  Výhoda: výpočtová nenáročnosť.

  Nevýhoda: nemusí zachytiť dôležitosť niektorých premenných.

- *Wrapper methods* - tieto metódy používajú na výber premenných prediktívne modely. Pre každý podsúbor dát (oklieštený o nejaký počet premenných) vypočíta predikčnú chybu na validačnej vzorke a na základe výsledkov vyberie tie premenné, ktoré mali chybu najnižšiu. Medzi tieto metódy patria napríklad rekurzívny výber premenných, postupný výber premenných, Hill Climbing, genetické algoritmy,...

  Výhoda: premenné sú určené predikčným modelom, ktorý vie zachytiť závislosť premenných, ktorú filtrovacia metóda nezachytí.

Nevýhoda: pri malej vzorke dát (v zmysle pozorovaní) majú tieto metódy tendenciu pretrénovania sa pri výbere premenných. Naopak pri veľkej vzorke dát (v zmysle počtu premenných) sú tieto modely veľmi výpočtovo náročné.

- *Embedded methods* - poslednou z uvedených metód sú takzvané vnorené metódy. Jedná sa o spôsob výberu premenných, ktorý kombinuje pozitívne vlastnosti oboch vyššie spomenutých metód. Výber premenných totiž prebieha zároveň s trénovaním klasifikačného modelu vďaka vlastnému výberovému procesu zabezpečeného regularizáciou. Hlavnými z uvedených metód sú Lasso regresia, Elastic net a rozhodovacie stromy, respektíve náhodný les.
  Výhoda: Simultánny výber premenných zároveň s trénovaním modelu.

Z každého okruhu sme vybrali niekoľko metód, ktoré boli následne aplikované na zvolený dátový súbor. Nižšie si uvedieme stručnú charakteristiku každej zo zvolenej metódy ako aj jej zaradenie.

- *Korelačná filtrovacia metóda* [12] je založená na pozorovaní korelačných koeficientov vstupných premenných. Predpoklad metódy je ten, že vysoko korelované vstupné premenné neprinášajú dodatočnú informáciu pre predikciu modelu a preto sú z ďalšieho procesu vynechané. *Cutoff rate*, teda hodnota korelačného koeficientu pri ktorom odstraňujeme premenné, bol nastavený na hodnotu 0.75.
  Na výpočet v softevéri R bola použitá funkcia *findCorrelation* z balíka *corrplot* [50].

- *Information gain* [41] je ďalšou z filtrovacích metód, ktorej hlavnou úlohou je určiť dôležitosť každej premennej vzhľadom na vysvetľovanú premennú pomocou entropie, teda miery neistoty. Tá počíta s podmienenou pravdepodobnosťou vysvetľujúcej premennej poznajúc hodnoty vysvetľovanej premennej. Keďže výstupom metódy je len hodnota veľkosti prínosu informácie jednotlivých premenných k výslednej premennej (to znamená, že nevyberá priamo konkrétne premenné), rozhodli sme sa vybrať tie premenné, ktorých hodnota *information gain* bola vyššia ako priemer.
  Na výpočet v softevéri R bola použitá funkcia *attrEval* z balíka *CORElearn* [38].

- *Statistically equivalent signature* [30] sa radí medzi algoritmy, ktoré sa zakladajú na teórii kauzálnej analýzy, teda analýzy zisťujúcej príčiny nastatia javov, čo v prípade

klasifikácie znamená zaradenie do danej triedy. Metóda na základe predpokladov tejto teórie dostáva optimálnu sadu premenných použitím Bayesovskej siete, čo je istá pravdepodobnostná grafická metóda. Úlohou Bayesovských sietí je modelovať podmienené závislosti premenných voči vysvetľovanej premennej. Deje sa tak pomocou ich reprezentácie vo forme hrán v orientovanom grafe. Tie sú nasledovne použité vo výberovom algoritme. Keďže v R-ku metóda priamo vyberá premenné, nebolo potrebné použiť žiadne kritérium na ich výber.
Na výpočet v softevéri R bola použitá funkcia *SES* z balíka *MXM* [30].

- *Selection by filter* [29] je poslednou z použitých filtrovacích metód. Algoritmus pristupuje postupne ku každej vysvetľujúcej premennej a na základe zaradenia do klasifikačných tried porovnáva priemery hodnôt danej premennej. Používa pritom ANOVA test, ktorý určuje štatistickú signifikantnosť rozdielu priemerov. Ak je rozdiel signifikantný, tak premennú zachováva, v opačnom prípade ju nezaradí do výberu premenných. Metóda určuje aj skóre jednotlivých premenných a to pomocou p-hodnoty štatistického testu.
  Na výpočet v softevéri R bola použitá funkcia *sbf* z balíka *caret* [28].

- *Recursive feature elimination* [29] je metóda z kategórie wrapper methods, ktorá sa snaží vybrať optimálny súbor premenných na základe úspešnosti výsledkov klasifikačného algoritmu (my sme vybrali random forest). Algoritmus najprv natrénuje model s použitím všetkých premenných a vypočíta dôležitosť každej z nich. V ďalšom kroku odstráni premennú, ktorá má najnižšiu dôležitosť. Opäť natrénuje model, znova určí dôležitosť premenných a opäť jednu odstráni. Tento proces sa opakuje niekoľko krát (vopred sa určí koľko) a následne sa porovná presnosť každého z vybraných súborov premenných. Finálne vyberáme tie premenné, pri ktorých model hlásil najvyššiu presnosť predikcie.
  Na výpočet v softevéri R bola použitá funkcia *rfe* z balíka *caret* [28].

- *Forward-backward selection* [3] je podobná predchádzajúcej metóde s rozdielom, že v prvom kroku sa začína s prázdnou množinou premenných a postupne sa premenné pridávajú podľa toho, ktorá z premenných najviac zvyšuje presnosť modelu (použitá bola kvadratická diskriminačná analýza). V druhom kroku metóda identifikuje

najhoršiu premennú z aktuálneho výberu a porovnáva presnosť modelu po jej odstránení s presnosťou modelu keď sa tam premenná nachádzala. Uvedeným spôsobom výberu premenných tak metóda prihliada aj na interakciu medzi aktuálne vybranými premennými (zabezpečené v druhom kroku).
Na výpočet v softevéri R bola použitá funkcia *stepclass* z balíka *klaR* [51].

- *LASSO* [11] metóda je v podstate logistickou regresiu s použitím regularizácie vo forme l-1 normy. Ako sme už spomínali, l-1 norma penalizuje rovnako silno všetky hodnoty β, ktoré predstavujú hodnoty koeficientov pri premenných v modely. V prípade, že niektorá z premenných prispieva do modelu malou váhou, LASSO na základe l-1 normy pošle takúto β do 0. Týmto spôsobom teda prebieha samotný výber premenných. LASSO predstavuje metódu z kategórie *embedded methods*.
Na výpočet v softevéri R bola použitá funkcia *cv.glmnet* z balíka *glmnet* [8].

- *Random forest selection* [11] metóda je druhou z kategórie *embedded methods* a zároveň poslednou z použitých metód na výber premenných. V každej z iterácii krížovej validácie sa model natrénuje pomocou klasifikačnej metódy random forest a určí tak dôležitosť premenných pomocou Giniho indexu. Ten meria takzvanú „nečistotu“ uzlov, teda miest, kde sa strom rozhoduje, po ktorej vetve bude ďalej pokračovať. Keďže Giniho index určuje dôležitosť jednotlivých premenných a nerobí ich finálny výber, zobrali sme len tie, ktoré mali vyššiu hodnotu dôležitosti ako priemer.
Na výpočet v softevéri R bola použitá funkcia *randomForest* z rovnomenného balíka *randomForest* [33].

5.2 Predstavenie dát

Pred uvedením výsledkov si predstavíme dátový súbor, ktorý bol na dané analýzy a porovnania zvolený. V predchádzajúcej kapitole sme pracovali s dátovým súborom s údajmi o ochorení diabetes. Ten však obsahoval malé množstvo vstupných premenných, ktoré už zároveň boli starostlivo vybrané a nebolo potrebné ich prerieďovať a vyťažovať tie najdôležitejšie z nich. Vybrali sme preto dáta spomenuté v článku [17], ktoré obsahujú údaje o finančných auditoch vykonaných v Indii na 776 spoločnostiach z 14 rôznych oblastí biznisu (súbor možno nájsť na stránke Kaggle [26]). Súbor obsahuje 27 premenných pričom jedna

z nich je vysvetľovaná premenná, ktorá informuje o tom, či zistenia auditu v jednotlivých spoločnostiach sú závažné a treba ich hlbšie prešetriť.

Tak ako v predchádzajúcom prípade aj teraz sme vysvetľovanú premennú pretransformovali na hodnoty $y \in \{-1, 1\}$. Uvádzame aj stručné zhrnutie významu najdôležitejších premenných, tak ako sa uvádza v [17]:

- *PARA_A* - zistená finančná nezrovnalosť v reporte s označením A (v 10 miliónoch rupií),
- *PARA_B* - zistená finančná nezrovnalosť v reporte s označením B (v 10 miliónoch rupií),
- *Number* - historické rizikové skóre nezrovnalosti finančných údajov,
- *MoneyValue* - hodnota finančných nezrovnalostí minulých auditov (v 10 miliónoch rupií),
- *DistrictScore* - historické rizikové skóre za posledných 10 rokov v určitej geografickej oblasti,
- *HistoryLoss* - počet rokov v strate za posledných 10 rokov,
- *Sector_score* - hodnota historického rizikového skóre sektora,
- *ARS* - kombinované auditné riziko získané analytickým procesom,
- *Inherent_Risk* - finančné riziko získané súčtom rizikových hodnôt premenných *PARA_A*, *PARA_B*, *Number* a *MoneyValue*,
- *CONTROL_RISK* - finančné riziko získané súčtom rizikových hodnôt premenných *DistrictScore* a *HistoryLoss*,
- *Detection_Risk* - riziková hodnota nezachytená auditom,
- *Audit_Risk* - výsledná hodnota auditného rizika získaná súčinom rizík *Inherent_Risk*, *CONTROL_RISK* a *Detection_Risk*. Na základe tejto premennej bola modelovaná hodnota vysvetľovanej premennej y.

Pre premenné *PARA_A*, *PARA_B*, *Number*, *MoneyValue*, *DistrictScore* a *HistoryLoss* sú uvedené aj pravdepodobnosti pochybenia ako aj výsledné rizikové skóre získané súčinom pravdepodobnosti a možnej finančnej straty.

Mimo týchto premenných boli v datasete aj premenné *LOCATION_ID* a *TOTAL*, ktoré sme však odstránili, pretože neobsahovali žiadnu užitočnú informáciu. Odstránené boli taktiež premenné *Inherent_Risk*, *CONTROL_RISK* a *Detection_Risk* z toho dôvodu, že boli získané lineárnou kombináciou iných stĺpcov ako aj premenná *Audit_Risk*, ktorá je v podstate pre nás pri predikcii neznáma.

Odstránené bolo aj jedno z pozorovaní, keďže obsahovalo NA hodnotu. Dostávame tak výslednú dátovú vzorku so 775 meraniami a 19 vysvetľujúcimi premennými.

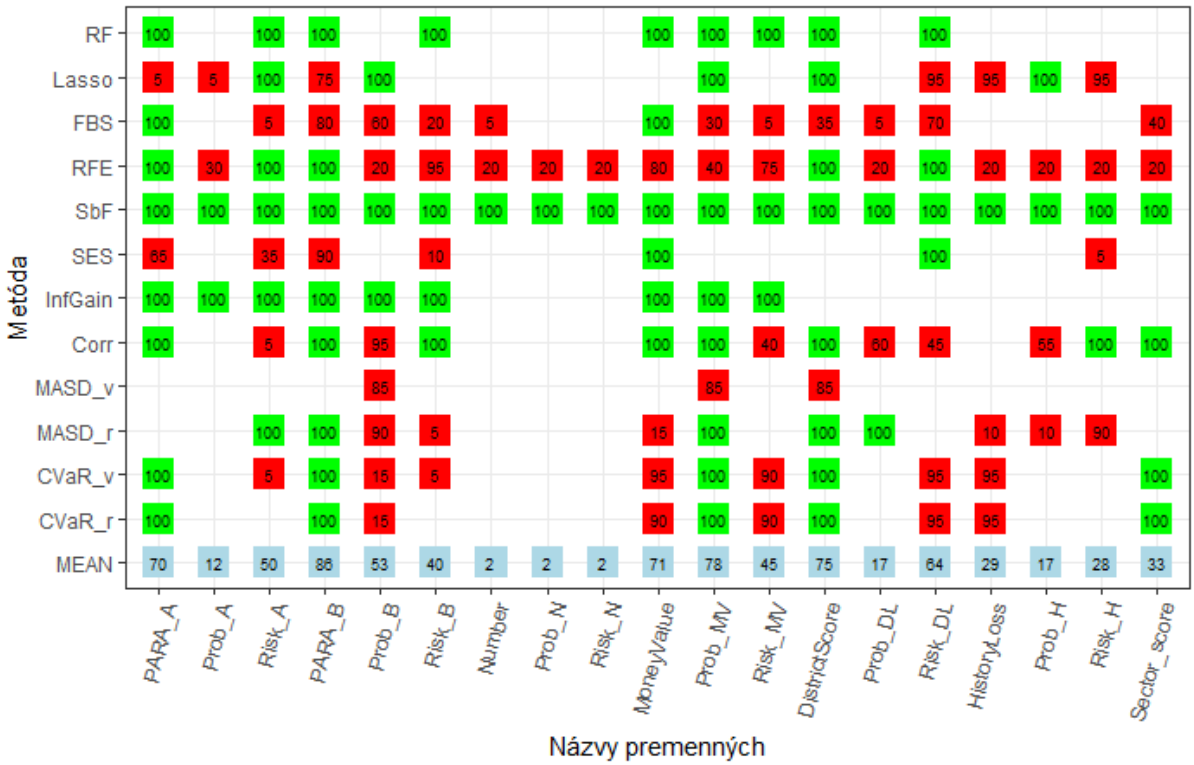
Takisto ako pri dátach o diabetese, aj tu sme štatisticky vyhodnocovali vplyv premenných na výslednú premennú. Jediný prípad, v ktorom sa štatistický test preukázal signifikantný bola premenná *Sector_score* a preto sme ju prenásobením záporným znamienkom. Takto transformovaná premenná potom vstupovala do modelov CVaR a MASD.

5.3 Porovnanie metód

Po týchto krokoch nám už nič nebránilo v tom, aby sme pristúpili k porovnávaniu samotných metód určených na výber premenných. Ako prvé sme pozorovali, ktoré z premenných zvolené metódy najčastejšie vyberajú. Opäť bola použitá krížová validácia, avšak v tomto prípade sme sa nesnažili nájsť optimálne hodnoty parametrov ako v sekcii 4.5 pri trénovaní modelov CVaR a MASD, ale sledovali sme koľko krát boli vybrané jednotlivé premenné vyššie popísanými metódami, pri použití $k = 20$ trénovacích zložiek určených na krížovú validáciu. Granularita možného výskytu premennej vo výsledkoch je teda 5%. Výsledky sú zhrnuté na Obrázku 12.

Zelenou farbou sú na Obrázku 12 vyznačené tie metódy, ktoré boli vybrané vo všetkých 20 iteráciách krížovej validácie, t.j. v 100% prípadov. Červenou farbou sú potom označené tie premenné, ktoré sa vyskytli v najviac 95% prípadov.

Môžeme pozorovať, že metóda *selection by filter* (SbF) nachádzajúca sa vo výslednej tabuľke, ani v jednom prípade nevyhodila žiadnu z premenných. Inými slovami, za každých okolností by algoritmus pracoval s celým pôvodným dátovým súborom. Z definície algoritmu SbF teda vyplýva, že rozdiely priemerov premenných v závislosti od zaradenia



Obr. 12: Percentuálne vyjadrenie výberu premenných po krížovej validácii.

do triedy sú signifikantné a preto žiadna z premenných nebola vyradená.

Čo sa týka algoritmov *random forest* (RF) a *information gain* (InfGain), tie na rozdiel od SbF vyhodili viaceré premenné z pôvodného datasetu, avšak v každej iterácii krížovej validácie ponechali tie isté premenné. Je teda zrejmé, že resamplovanie trénovacej vzorky dát v procese krížovej validácie nijakým spôsobom nevplýva pri uvedených dvoch algoritmoch na výber premenných.

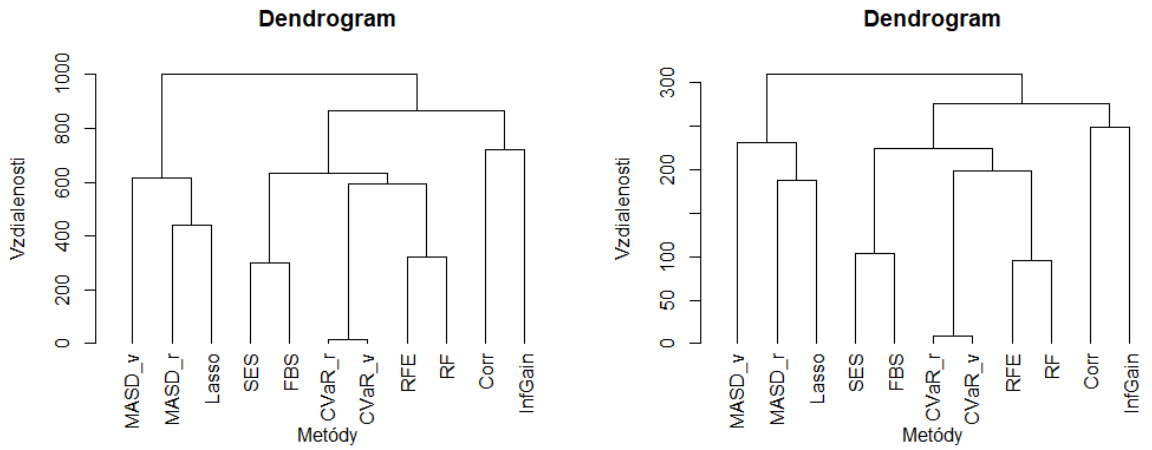
Navrhované metódy CVaR_r a CVaR_v taktiež javia známky konzistentnosti vo výbere parametrov. Takmer každá z premenných, ktorú určili metódy CVaR_r a CVaR_v je zvolená minimálne v 90% prípadov, až na premenné *Risk_A, Prob_B* a *Risk_B*, ktoré boli vybrané najviac v 15% pre obe metódy. Riadok s označením MEAN predstavuje priemerný percentuálny výber premenných naprieč všetkými metódami (okrem SbF). Pri porovnávaní týchto hodnôt s hodnotami premenných našich metód CVaR_r a CVaR_v je zaujímavé pozorovať, že každá z premenných, ktorá bola vybraná vo viac ako 50% prípadov riadku MEAN (dokopy 7), bola vybraná aj metódami CVaR_r a CVaR_v. Z tohto pozorovania môžeme konštatovať, že metódy CVaR_r a CVaR_v vyberali najmä také premenné, ktoré považovali aj ostatné metódy za najdôležitejšie, respektíve za najväčších nositeľov infor-

mácie. Metódy CVaR vybrali aj premennú *Sector_score*, ktorá bola v predprípravе dát prenásobená záporným znamienkom. Pre zaujímavosť sme metódu nechali zbehnúť aj pre netransformovaný dataset, v ktorej výsledkoch sa však táto premenná nenachádzala. Môžeme preto predpokladať, že napriek nedostatkom štatistického testovania uvedeného v časti 4.2.2, bolo dobré daný stĺpec transformovať.

Metóda MASD_r vybrala 11 premenných, z toho až 7 z nich s výskytom v aspoň 90% prípadov krížovej validácie. Naopak zvyšné 4 sa vyskytli v najviac 15% prípadov. Zaujímavou je metóda MASD_v, ktorá vybrala len 3 premenné spomedzi všetkých a aj to nie vo všetkých prípadoch (to znamená, že v niektorých prípadoch vyberala len dve alebo dokonca len 1 premennú). Tieto tri sú však vo veľkej miere vyberané aj metódou MASD_r.

Pomerne veľká nekonzistentnosť sa ukazuje pri metódach FBS a RFE. Aj tieto metódy majú málo takých premenných, ktoré vyberali s vysokou frekvenciou a veľa tých, ktoré vybrali len niekoľko krát. Tento fakt navyše podporuje skutočnosť, že metóda RFE vybrala počas krížovej validácie každú z premenných minimálne raz.

Lasso a Corr vyberajú pomerne veľké množstvo premenných s pomerne vysokou frekvenciou, kdežto metóda SES vybrala počas všetkých epoch krížovej validácie len 7 premenných a len 4 z nich vo viac ako 50% prípadov.



(a) Použitie l-1 normy.

(b) Použitie l-2 normy.

Obr. 13: Znázornenie proximity metód výberu premenných pomocou dendrogramov za použitia dvoch rôznych metrík vzdialenosti.

Pre lepšie pochopenie vzťahov medzi vektormi frekvencie vybraných premenných jed-

notlivými metódami počas krížovej validácie sme vytvorili maticu vzdialeností týchto vektorov a tú sme sa pokúsili premietnuť do dendrogramu. Na Obrázku 13 ponúkame náhľad na dva typy vyobrazení s použitím l-1 a l-2 noriem, znázorňujúce „blízkosť“, respektíve „vzdialenosť“ metód na výber premenných. Ypsilonová os nám ukazuje ako sú vektory od seba vzdialené v jednotlivých normách.

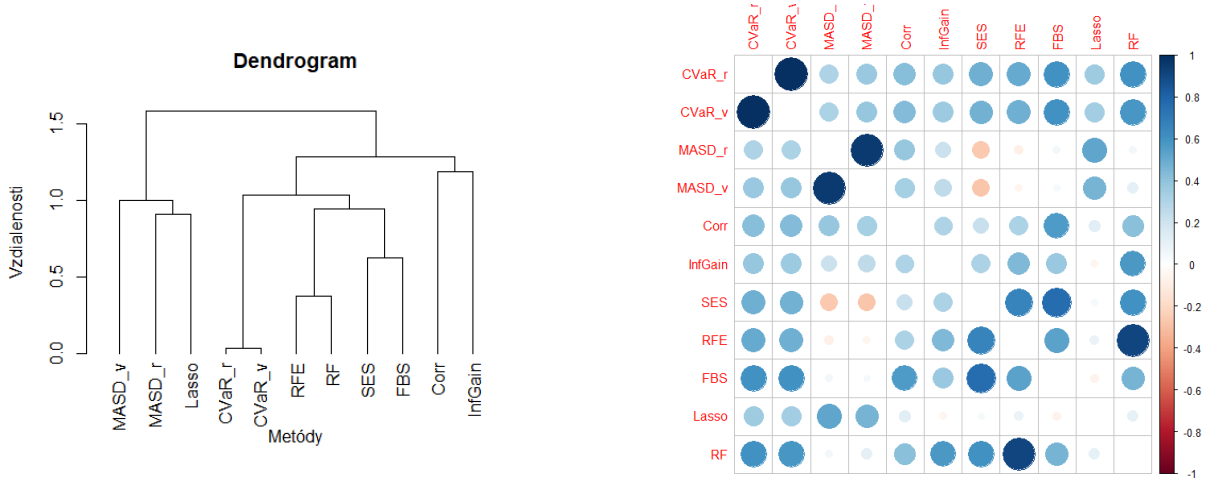
Pri použití oboch noriem vidíme, že metódy CVaR_r a CVaR_v majú k sebe najbližšie spomedzi všetkých porovnávaných metód. Je to pomerne očakávaný jav, keďže sa jedná o tú istú metódu s rozdielom použitia rozličnej pravdepodobnosti. Napriek tomu, že metódy MASD_r a MASD_v sa líšia len v tom ako majú počítanú pravdepodobnosť v modely, tak ako metódy CVaR, vzájomná vzdialenosť ich vektorov je podstatne väčšia a to za použitia oboch noriem.

Očakávali by sme, že ostatné porovnávané metódy sa budú zoskupovať podľa zaradenia do okruhu podobných metód (filter, wrapper a embedded methods) avšak treba si uvedomiť, že jedná sa len o prevedenie akým premenné vyberajú. Zaujímavejšie je pozrieť sa na to, aké algoritmy sa pri procese výberu používajú.

Metódy RF a RFE sú relatívne blízko seba napriek tomu, že prvá spomínaná je z okruhu embedded methods a druhá z filter methods. Rozhodujúce však je, že obe na výber premenných používajú algoritmus random forest, ktorý počíta s dôležitosťou premenných.

Ďalšou z vybraných dvojíc sú metódy SES a FBS, kde obe na výber premenných používajú Bayesovskú štatistiku. Zvyšné z metód sú od seba už relatívne dosť vzdialené, preto nemá zmysel hľadať hlbší súvis ich spoločného zlúčenia ako ten, že obe sú rovnako ďaleko od zvyšných metód.

Posledná z vecí, na ktoré sa v našej analýze pozrieme sú korelácie vybraných premenných spomedzi zvolených metód. Prvý z obrázkov, Obrázok 14a, opäť zobrazuje dendrogram vzdialeností, tentokrát však počítaný zo vzájomných korelácií. Na výpočet vzdialeností za použitia korelácie bol použitý vzorec $d = \sqrt{2 * (1 - \text{korelácia})}$, kde korelácia značí koreláciu dvojice vektorov. Výsledky sú však veľmi podobné ako na Obrázku 13. Spojením informácie vyobrazených Obrázkov 14a a 14b a poznatkov z predchádzajúcich analýz sa metódy MASD spolu s Lassom javia, že vyberajú podstatne rozličné premenné ako zvyšné metódy určené na výber premenných. Naopak metódy CVaR sa javia, že vyberajú veľmi podobné premenné ako zvyšné metódy.



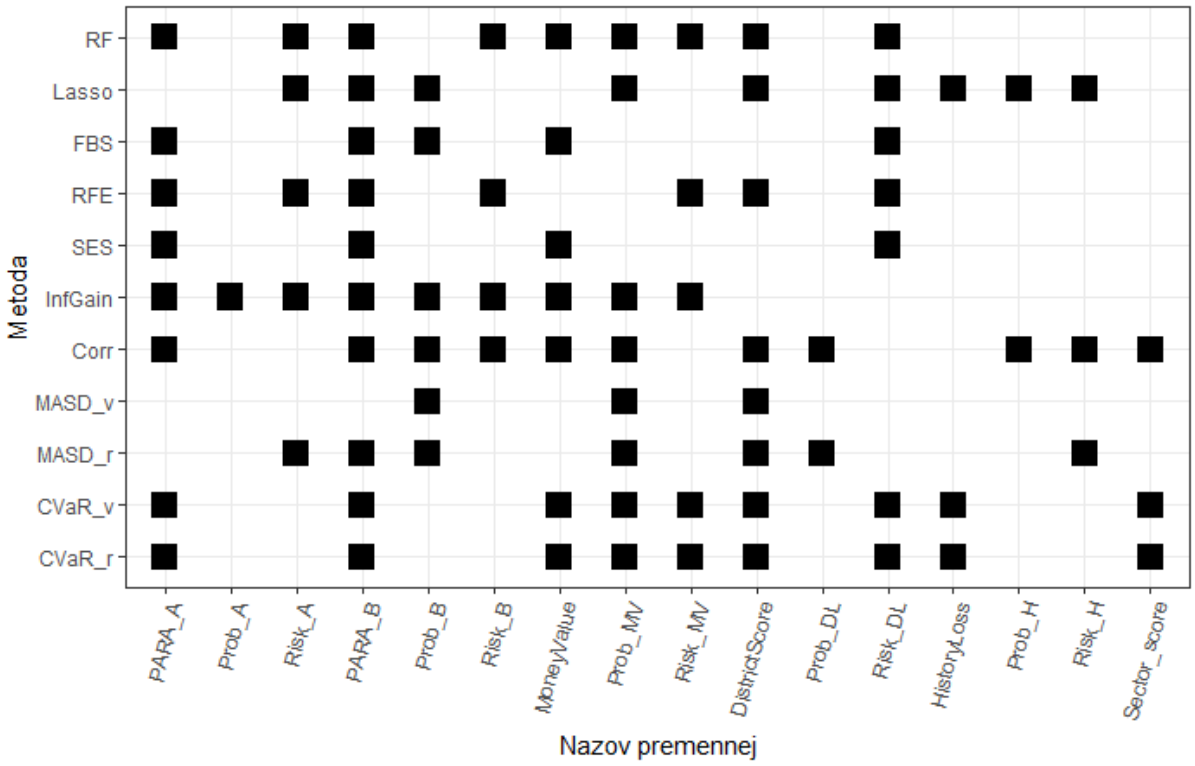
(a) Graf zobrazujúci vzájomné korelácie metód.

(b) Tabuľka korelácií vektorov vybraných premenných.

Obr. 14: Znázornenie korelácii metód výberu premenných.

5.4 Aplikácia v logistickej regresii

V predchádzajúcej podkapitole sme si uviedli viaceré existujúce metódy určené na výber premenných a ich výsledky sme porovnávali a analyzovali s možnou aplikáciou metód



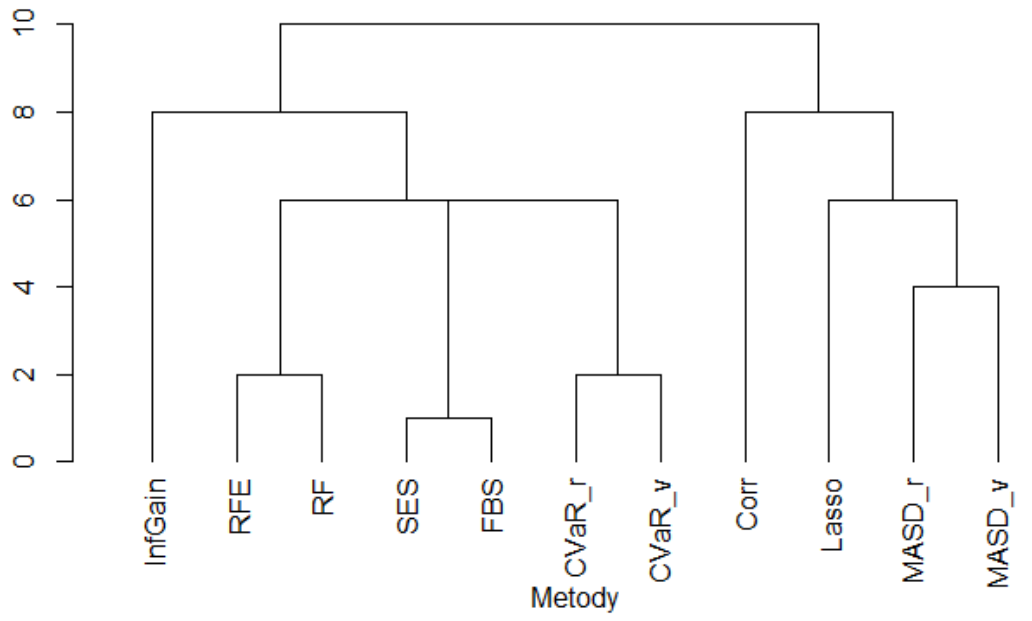
Obr. 15: Vizualizácia výberu premenných.

CVaR a MASD v oblasti výberu premenných. O správnosti výberu premenných z podkapitoly 5.3 sa viac dozvieme v tejto časti, kde budeme porovnávať metriky úspešnosti na logistickej regresii v závislosti od toho, aké vstupné premenné nám odporučia metódy na výber premenných.

Všetky zo spomenutých metód určených na výber premenných sme opäť spustili, tentokrát na celej trénovacej vzorke bez krížovej validácie. Na Obrázku 15 môžeme pozorovať, ktoré premenné metódy finálne vybrali na predikciu. Premenné *Number, Prob_N* a *Risk_N*, ktoré neboli vybrané ani jednou metódou, sa na obrázku nenachádzajú. Z tabuľky bola taktiež vynechaná metóda SbF, ktorá zvolila všetky premenné.

Obe metódy CVaR spoločne vybrali 9 rovnakých premenných. Rozdiel medzi výberom z predchádzajúcej sekcie 5.3 na Obrázku 12 je vynechanie tých premenných, ktoré boli zvolené len v malom počte prípadov. Tak isto sa zachovala aj metóda MASD_r, ktorá ponechala len premenné s najčastejším výberom.

Podobne sa zachovali aj ostatné metódy a vybrali len tie premenné, ktoré sa vyskytovali vo viac ako 60% prípadov. Metóda RFE dokonca nevybrala jednu premennú, ktorú v predchádzajúcej časti vybrala v 80% prípadoch krížovou validáciou.



Obr. 16: Dendrogram znázorňujúci počet odlišne vybraných premenných naprieč metódami.

Podobnosť výberu premenných možno vidieť na Obrázku 16, ktorý znázorňuje počet rozdielnych premenných zvolených jednotlivými metódami.

Keďže metóda logistickej regresie s l-2 regularizáciou nemá vlastnosť výberu premenných, snažili sme sa porovnať, aké výsledky bude dosahovať LR, ak priestor vstupných premenných trénovacieho a testovacieho datasetu znížime vzhľadom na výsledky výberu premenných uvedených na Obrázku 15. Tabuľka 7 nám poskytuje porovnanie výsledkov vzhľadom na tri zvolené metriky.

	ACC	**Sensitivity**	**Specificity**	**# premenných**
LR	0.9290	**1.0000**	0.8254	19
CVaR_r	0.9226	0.9891	0.8254	9
CVaR_v	0.9226	0.9891	0.8254	9
MASD_r	**0.9484**	0.9674	**0.9206**	7
MASD_v	**0.9484**	0.9674	**0.9206**	**3**
Corr	**0.9484**	0.9891	0.8889	11
InfGain	0.8516	1.0000	0.6349	9
SES	0.5935	1.0000	0.0000	4
RFE	0.6516	0.9130	0.2698	7
FBS	0.8065	0.9674	0.5714	5
Lasso	**0.9484**	0.9674	0.9206	9
RF	0.5935	1.0000	0.0000	9

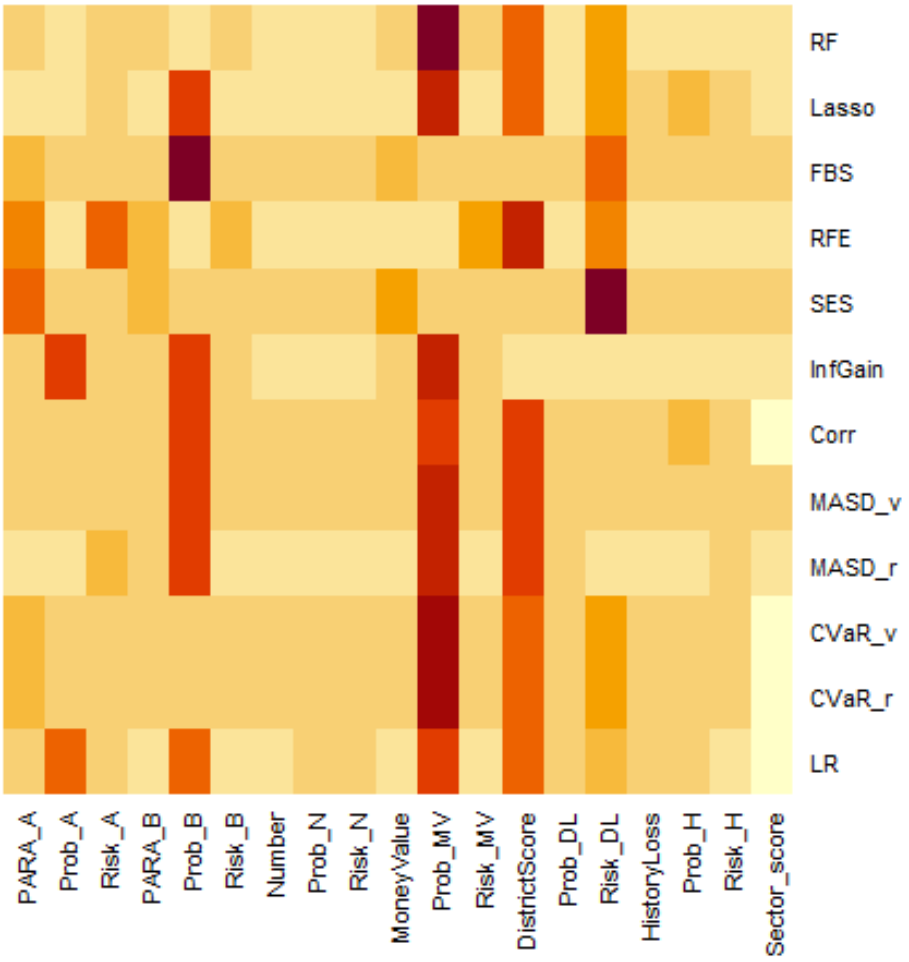
Tabuľka 7: Výsledky predikcie logistickej regresie pre rôzny výber vstupných premenných.

Dostávame zaujímavé výsledky, kde až v štyroch prípadoch pozorujeme, že logistická regresia s okresaným datasetom dosahuje lepšie výsledky v metrike ACC ako za použitia všetkých pôvodných premenných, tak ako to robí LR. Navyše sa tak deje pri troch metódach, MASD_r, MASD_v a Lasso, ktoré sa svojimi výsledkami v predchádzajúcej sekcii 5.3 najväčšmi líšili od výberu premenných ostatnými metódami. Okrem nich tieto výsledky dosahuje aj metóda Corr, ktorá svojimi výsledkami bola taktiež veľmi odlišná od ostatných metód. Ak do porovnania dáme aj počet použitých premenných na výpočet úspešnosti metódy, tak jednoznačne z porovnaní vychádza najlepšie LR s vybranými premennými metódou MASD_v, ktorá na výpočet predikcie použila len tri premenné (*Prob_B,*

Prob_MV a *DistrictScore*). Hneď za ňou by sa umiestnila MASD_r, ktorá vybrala 7 premenných.

Dobré výsledky môžeme pozorovať aj pri metrikách LR s použitím premenných z metód CVaR_r a CVaR_v aj napriek tomu, že sú mierne horšie ako pri LR so všetkými premennými. Výhodou je, že na výpočet sa použije len 9 premenných namiesto 19. Pri veľkom dátovom súbore to tak vie značne ušetriť čas výpočtu.

Naopak veľmi zlé výsledky dosiahla LR s použitím premenných, ktoré pokladali za dôležité metódy RF, SES a RFE. Navyše pri prvých dvoch spomenutých metódach vidíme, že s danými vybranými premennými najlepšiu predikciu, ktorú dokáže LR dať je zaradenie pozorovania do kategórie kedy zistenie auditu nie je závažné a firmu netreba ďalej kontrolovať. Takýto prístup je samozrejme nesprávny.



Obr. 17: Váhy premenných naprieč výsledkami LR.

Na Obrázku 17 ponúkame ešte jednu z vizualizácií výsledkov, pričom obrázok treba

čítať po riadkoch. Najbledšie hodnoty vrámci riadku znamenajú nulovú váhu premennej v tom danom modeli, naopak tmavé vyjadrujú, že daná metóda jej priradila vysokú váhu voči iným premenným. Z obrázku sa dá vyčítať, že najvyššiu váhu vrámci metód dostali premenné *Prob_B, Prob_MV* a *DistrictScore*. Práve tieto premenné vybrala každá z metód, ktorá v porovnaniach predčila LR s celým pôvodným datasetom.

Po zanalyzovaní výsledkov môžeme potvrdiť, že robiť riadnu prípravu dát predtým, ako ideme klasifikovať nové dáta, vo forme výberu premenných sa javí ako účinný prostriedok na zlepšenie predikcií. Navyše medzi veľmi účinné metódy výberu premenných pre logistickú regresiu sa javia práve skúmané metódy CVaR a MASD.

6 Aplikácia metód CVaR a MASD na dáta z oblasti energetiky

V poslednej kapitole urobíme kompletnú analýzu a porovnanie metód CVaR a MASD s ostatnými štatistickými metódami uvedenými v časti 4.4 na dátach o predikovaní výroby elektrickej energie (vyrábať/nevyrábať) plynovými elektrárňami na Slovensku. Tie majú vlastnosť, že sú veľmi flexibilné a ľahko vedia zapnúť a vypnúť svoju výrobu, čo by nebolo možné napríklad pri jadrovým elektrárňach, ktoré musia vyrábať stále alebo veterných elektrárňach, ktoré vyrábajú podľa poveternostných podmienok. V bežnej praxi sa predikuje cena elektrickej energie a pomocou nej sa určuje, či elektráreň bude vyrábať alebo nie. My sme sa však rozhodli zvoliť priamy prístup a teda priamo z dát predikovať výrobu bez medzi kroku výpočtu ceny elektrickej energie.

6.1 Predstavenie a analýza dát

Keďže európsky trh s energiami je veľmi úzko prepojený a sú na ňom viacerí silní hráči ako napríklad Nemecko a Francúzsko, na predikciu budeme používať predovšetkým dáta z týchto krajín. Dátový súbor obsahuje nasledovné premenné:

- Spot.EEX.EEX.FIX - hodinové údaje cien elektriny (z [45]),
- CO2 - cena CO_2 povoleniek,
- GAS - cena plynu,
- COAL - cena uhlia (všetky komodity získané z [22]),
- CONS - spotreba elektrickej energie (v MW) krajín Nemecka, Rakúska, Francúzska, Poľska, Česka, Belgicka, Holandska a Dánska,
- WON a WOF - výroba pevninovej a mimo pevninovej veternej energie spomínaných krajín,
- SOL - výroba solárnej energie spomínaných krajín,
- NUC - výroba jadrovej energie spomínaných krajín,

- HWR a ROR - výroba vodnej energie spomínaných krajín (spotreby a výroby získané zo stránky [7]),
- Ten a Hz - maximálne kapacity predanej elektriny z Nemecka do Čiech a Poľska (zo stránky [23]).

Za pozorované obdobie sme si zobrali dáta za rok 2019. Okrem dát získaných z webových stránok sme do datasetu doplnili informáciu o tom, či je daný deň pracovný ako aj informáciu o aktuálnom mesiaci.

Predikovanú premennú $y \in \{-1, 1\}$, sme určovali na základe rozdielu ceny elektrickej energie a ukazovateľa Short-Run Marginal Costs (SRMC), ktorý určuje hraničné náklady, pri ktorých sa oplatí elektrárni vyrábať. Rozdiel týchto dvoch cien sa zvykne nazývať *clean spark spread* [40]. Ak bol rozdiel týchto dvoch cien väčší ako 0, zvolili sme $y = 1$, teda oplatí sa vyrábať. V opačnom prípade sme zvolili $y = -1$. Ukazovateľ SRMC je možné vypočítať zo získaných dát nasledovným vzorcom:

$$\text{SRMC} = \frac{\text{Cena plynu}}{\text{Efektivita}} + \text{Cena } CO_2 \text{ povoleniek} * \frac{\text{Intenzita emisií} * \text{Energetická konverzia}}{\text{Efektivita}},$$

kde cenu plynu poznáme z dát, efektivitu využitia plynu na výrobu elektrickej energie sme zvolili 55%, cenu CO_2 povoleniek poznáme taktiež z dát a posledný zlomok má konštantnú hodnotu po zaokrúhlení 0.33 (získané z [40]).

Pri rozhodovaní elektrárne, či na ďalší deň vyrábať alebo nie, sú kľúčové najmä cena elektriny, plynu a CO_2 povoleniek. Cena plynu a CO_2 povoleniek sú náklady, ktoré sú známe, pretože sa určujú na deň vopred. To znamená, že v čase rozhodovania elektráreň vie, aké budú jej náklady na výrobu (hodnota SRMC). Čo ale v tom čase nepozná, je cena elektriny. O nej priamo rozhoduje výška odhadovanej spotreby odberateľov elektrickej energie, ktorá je síce taktiež len estimáciou, avšak veľmi presnou, pričom z dlhodobého hľadiska sa jej priemer blíži k tomu skutočnému.

Keďže cena elektriny je veličina, ktorú dopredu nepoznáme a bola potrebná len na výpočet predikovanej premennej, nezahŕňame ju ďalej do nášho datasetu. Z dôvodu výpočtu SRMC z ceny plynu a ceny CO_2 povoleniek boli tieto premenné lineárne závislé, preto sme sa rozhodli vynechať premennú ceny plynu. Finálny dátový súbor tak obsahuje 50 premenných (z toho jednu predikovanú) s 8707 pozorovaniami.

Premenné	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**	**Transf.**
SRMC	21.39	27.72	32.16	32.82	36.15	49.13	áno
CO2	19.20	23.99	25.32	25.20	26.56	30.20	nie
COAL	42.38	50.70	53.73	55.40	57.11	75.26	áno
CONS_DE	34335	47660	55415	55787	64216	77633	nie
CONS_AT	4176	6010	7130	7146	8135	10779	nie
CONS_FR	31183	44906	51818	53346	60702	86341	nie
CONS_PL	11400	16599	19419	19283	21831	26136	nie
CONS_CZ	4471	6562	7539	7553	8453	10901	nie
CONS_BE	0	8631	9710	9688	10631	13616	nie
CONS_NL	0	10046	11500	11726	13071	18574	nie
CONS_DK	1693	3244	3786	3828	4401	6750	nie
WON_DE	269	4774	8795	11360	15540	40225	áno
WON_AT	4	198	643	908	1505	2967	áno
WON_FR	521	1696	2845	3735	5140	12742	áno
WON_DK	9	467	920	1177	1745	4046	áno
WOF_DE	0	1210	2633	2758	4233	6825	áno
WOF_DK	0	282	596	641	978	3635	áno
SOL_DE	0	0	146	4788	7658	30028	nie
SOL_FR	0	0	182	1304	2346	6707	nie
SOL_CZ	0	0	18	257	410	1680	nie
SOL_AT	0	0	26	153	246	976	nie
NUC_DE	3382	7442	8607	8108	9230	9524	nie
NUC_FR	26139	38977	41541	43128	46534	56230	áno
NUC_CZ	1883	2912	3416	3262	3854	3972	áno
NUC_BE	2376	4085	4902	4728	5085	5886	nie
HWR_DE	1	77	129	151	209	769	nie
HWR_FR	0	815	1309	1574	2077	5894	nie
HWR_CZ	1	23	72	122	178	700	nie
HWR_AT	15	251	502	559	820	1631	nie

ROR_DE	1060	1450	1663	1658	1856	2176	nie
ROR_FR	1526	3581	4629	4608	5704	7744	nie
ROR_CZ	0	67	87	96	114	214	áno
ROR_AT	1047	2362	2956	3065	3725	5228	nie
Ten_CZ	0	407	949	833	1245	1519	áno
Hz_CZ	0	136	570	568	918	1583	áno
Hz_PL	0	60	224	253	387	1265	nie
WD[4]	-	-	-	-	-	-	nie
M_apríl	-	-	-	-	-	-	áno
M_august	-	-	-	-	-	-	nie
M_december	-	-	-	-	-	-	áno
M_február	-	-	-	-	-	-	áno
M_január	-	-	-	-	-	-	áno
M_júl	-	-	-	-	-	-	nie
M_jún	-	-	-	-	-	-	nie
M_máj	-	-	-	-	-	-	nie
M_marec	-	-	-	-	-	-	áno
M_november	-	-	-	-	-	-	nie
M_október	-	-	-	-	-	-	nie
M_september	-	-	-	-	-	-	nie

Tabuľka 8: Sumárna štatistika a prehľad dát.

Náhľad na dáta vo forme sumárnej štatistiky ponúkame v Tabuľke 8. Horizontálnymi čiarami sú oddelené spoločné skupiny premenných akými sú náklady na výrobu elektriny plynovými elektrárňami, ceny komodít, spotreba, výroba veterných, solárnych, jadrových a vodných elektrární (dva typy), maximálne kapacity prenosu a časové údaje.

Okrem úvodného čistenia dát, bolo opäť potrebné zistiť vplyv premenných na výslednú premennú y, tak ako tomu bolo v časti 4.2.2. Po aplikovaní štatistických testov na dáta zisťujeme, že analýza nám odporúča zmeniť znamienko pri 18 premenných. Pre premennú

[4]WD spolu s jednotlivými mesiacmi sú dummy premenné, preto neuvádzame ich sumárnu štatistiku. V Tabuľke 8 sa nachádzajú len z dôvodu informácie o transformácii premennej.

to znamená, že negatívne vplýva na predikovanú hodnotu y. Ktoré to sú je možné vidieť v Tabuľke 8 pod stĺpcom **Transf.**. Keďže sme si povedali, že toto testovanie nie je veľmi dôveryhodné, pretože nezachytáva možné kolinearity v dátach, je potrebné sa na transformované dáta pozrieť a kriticky zhodnotiť, či transformácie dávajú zmysel.

Z testovania premennej y vyplýva, že premenná SRMC má na ňu s rastúcimi hodnotami negatívny vplyv. Napriek neistote výsledkov testu zmenu akceptujeme. Totiž v prípade nemennej ceny elektriny a zároveň rastúcej hodnoty SRMC je zrejmé, že elektráreň nebude vyrábať, pretože by mala vyššie náklady ako by zarobila na predaji elektriny.

Pomerne prekvapivé je, že test neurčil za signifikantnú zmenu znamienka pri premennej CO2 povoleniek, ktorá priamo úmerne zvyšuje hodnotu SRMC. Usudzujeme však, že to môže byť z toho dôvodu, že od ceny CO2 povoleniek sa odvíja aj cena elektrickej energie, ktorá vstupuje do výpočtu y. Je ťažké určiť, či vplyv tejto premennej je väčší na SRMC alebo na cenu elektriny a z toho dôvodu nevieme logicky povedať, či by vplyv na y mal byť prirodzene pozitívny alebo negatívny. Z toho dôvodu nebudeme zasahovať do výsledkov testu.

Ďalším zistením štatistických testov je negatívny vplyv produkcie veterných elektrární na výrobu plynovou elektrárňou. Ako už bolo naznačené v úvode kapitoly, dôvodom môže byť nestálosť výroby veterných elektrární. Ak sa predpokladá, že nasledujúci deň má byť bezvetrie, t.j. veterné elektrárne nebudú môcť vyrábať, zníži sa celková produkcia energie avšak za predpokladu rovnakého dopytu. Z dôvodu nedostatku elektriny na trhu cena stúpne a plynové elektrárne začnú vyrábať (so ziskom) ako komplement k veterným elektrárňam. Výsledky testovania sa preto zdajú byť logické.

Pri premenných *NUC_FR, NUC_CZ, ROR_CZ, Ten_CZ* a *Hz_CZ* je už interpretácia zložitejšia ako v predchádzajúcich prípadoch, najmä z dôvodu, že znamienko bolo zmenené len niektorým premenným z rovnakej skupiny premenných. Preto do výsledkov taktiež nebudeme zasahovať.

Zmena znamienka nastala aj pre niektoré z mesiacov. Čo je zaujímavé, udialo sa tak práve pri chladných mesiacoch, t.j. od decembra až do apríla, pričom by sme očakávali presný opak z dôvodu vyššej spotreby energie. Tá sa však pri týchto testoch nezohľadňuje a preto len môžeme konštatovať, že vplyv chladnejších mesiacov, je na výroby negatívny.

6.2 Výsledky a porovnania

Na porovnania boli použité štatistické metódy a metódy strojového učenia uvedené v 3 a porovnávacie kritéria z podkapitoly 4.3. Trénovanie metód prebehlo na vzorke 6966 pozorovaní a testovanie na zvyšných 1741 pozorovaniach. Pri testovaní použili obe metódy CVaR hodnotu parametra $\beta = 0.75$ a taktiež obe metódy MASD hodnotu parametra $\tau = 0.8$. Neurónová sieť (NN) počítala s dvoma skrytými vrstvami, ktoré obsahovali 8 a 3 neuróny a ako aktivačnú funkciu sme zvolili hyperbolický tangens. Metóda kNN po krížovej validácii vybrala 4 susedov, podľa ktorých určovala kategórie. V prípade, že nastala zhoda a dvaja susedia boli z jednej a dvaja z druhej triedy, metóda rozhodla náhodne o tom, do ktorej triedy meranie zaradí. Pri zvyšných metódach nebolo potrebné určovať žiadne hyper-parametre ani dodatočné nastavenia, preto sme im nechali ich východzie nastavenia tak, ako sú uvedené v ich R-kovej dokumentácii. Výsledky metód predikujúcich spúšťanie výroby plynových elektrární môžeme sledovať v Tabuľke 9.

	Accuracy	Sensitivity	Specificity
CVaR_r	0.8903	0.7638	0.9371
CVaR_v	0.8742	0.8979	0.8655
MASD_r	0.7737	0.9553	0.7065
MASD_v	0.7588	**0.9596**	0.6845
LR	0.8771	0.6362	0.9662
SVM	0.9104	0.8000	0.9512
LDA	0.8966	0.7468	0.9520
NB	0.8179	0.8553	0.8041
NN	0.8943	0.6957	**0.9677**
DT	0.8604	0.7106	0.9158
RF	**0.9265**	0.8255	0.9638
kNN	0.9138	0.8234	0.9473

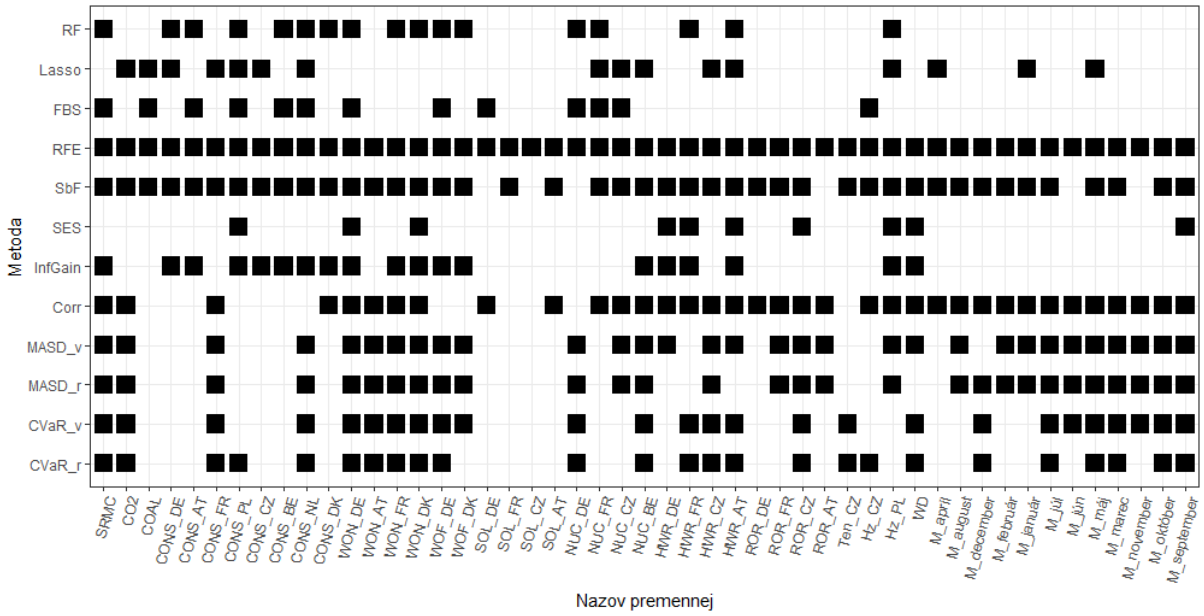
Tabuľka 9: Výsledky predikcie dát ohľadom výroby elektrickej energie.

Z Tabuľky 9 jasne vyplýva, že najúspešnejšie zaraďovala pozorovania do tried metóda RF. Tá si viedla dobre aj v metrike Specificity, najlepšie výsledky v nej však dosahuje metóda NN. Veľmi dobré výsledky v metrike ACC dosahujú taktiež metódy kNN a SVM.

Okrem metódy CVaR_r pozorujeme pri zvyšných metódach CVaR_v, MASD_r a MASD_v veľmi slabý výkon v metrike ACC. Naopak v metrike Sensitivity dosahujú tieto tri metódy najvyššie výsledky. Pri predikciách je preto kľúčové si uvedomiť, minimalizácia ktorej chyby je pre nás dôležitá. Pri našich dátach je však také niečo ťažké určiť, keďže nevieme povedať, aké náklady pri plynových elektrárňach sú spojené so zlou predikciou ohľadom výroby.

6.3 Výber premenných

Keďže energetický priemysel, tak ako aj finančný, je vysoko regulovaný, nie je možné použiť ľubovolný model na predikciu výroby alebo cien elektriny. Pre manažment podniku je potrebný jednoduchý a ľahko interpretovateľný model. Preto sa v mnohých prípadoch na predikcie používa obyčajná lineárna alebo logistická regresia. V predchádzajúcej časti 6.2 vyšli výsledky logistickej regresie vzhľadom na ostatné metódy pomerne dobre. V tejto podkapitole sa pozrieme, či použitím metód na výber premenných vieme výsledky LR zlepšiť alebo aspoň model dostatočne zjednodušiť. Postupujeme podobne ako v časti 5.4 z predchádzajúcej kapitoly 5.



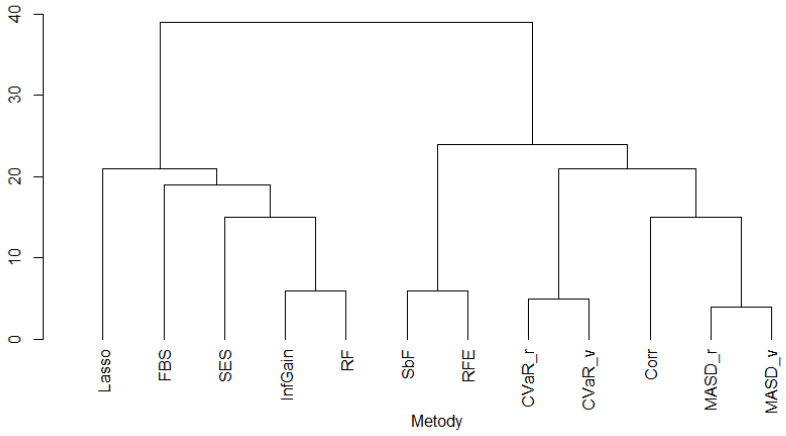
Obr. 18: Vizualizácia výberu premenných na dátach z energetiky.

Na Obrázku 18 pozorujeme, ktoré premenné boli zvolené metódami na výber premenných. Vysokú podobnosť výberov badáme pre metódy CVaR aj MASD. Pomerne podobné

premenné vybrala aj metóda Corr. Blízkosť badať aj pre dvojicu metód InfGain a RF.

RFE vybrala všetky premenné a tým sa pre nás stáva bezcennou, preto ju z ďalších porovnaní vynecháme. Metóda SbF vynechala do ďalšej fázy analýzy 6 premenných. Naopak metóda SES bola veľmi *sparse* a vybrala len malý počet premenných v počte 10. Pre zvyšné metódy nepozorujeme žiadnu výraznejšiu podobnosť alebo inú zaujímavosť.

Ponúkame ešte náhľad na Obrázok 19, ktorý nám opäť hovorí o blízkosti metód. Ten má na ypsilonovej osi uvedený počet rozdielnych vybraných premenných jednotlivými metódami. Potvrdilo sa, čo už bolo vidno z Obrázku 18 a teda podobnosť metód CVaR_r a CVaR_v, MASD_r a MASD_v spolu s Corr, ďalej metódy RF a InfGain, ako aj metódy RFE a Sbf, ktoré obe vybrali skoro každú z premenných.



Obr. 19: Dendrogram znázorňujúci počet odlišne vybraných premenných naprieč metódami pre dáta z oblasti energetiky.

Pomocou premenných vybraných z Obrázku 18 sme natrénovali niekoľko modelov logistickej regresie a následne otestovali na testovacích dátach. Výsledky porovnania sa nachádzajú Tabuľke 10.

Z Tabuľky 10 pozorujeme, že ani jedna z metód výberu premenných nevylepšila hodnotu metriky ACC logistickej regresie (s použitím všetkých premenných). Napriek tomu by sme výsledky úplne nezavrhovali. Metóda CVaR_r totiž dáva podobne dobré výsledky v metrike ACC, pričom používa len polovicu pôvodných premenných. Kombinácia výberu premenných pomocou metódy CVaR_r spolu s logistickou regresiou by preto mohla byť vhodnou alternatívou voči LR robenej na celom datasete. Vznikol by tak model, ktorý

	ACC	**Sensitivity**	**Specificity**	**# premenných**
LR	**0.8771**	**0.6362**	0.9662	49
CVaR_r	0.8621	0.5745	0.9685	25
CVaR_v	0.8530	0.5447	0.9670	26
MASD_r	0.8420	0.4979	0.9693	29
MASD_v	0.8518	0.5574	0.9607	31
Corr	0.8604	0.5660	0.9693	36
InfGain	0.8478	0.5638	0.9528	19
SES	0.8409	0.5234	0.9583	**10**
SbF	0.8708	0.6191	0.9638	43
FBS	0.8558	0.5234	**0.9788**	13
Lasso	0.8007	0.3979	0.9496	16
RF	0.8512	0.5638	0.9575	17

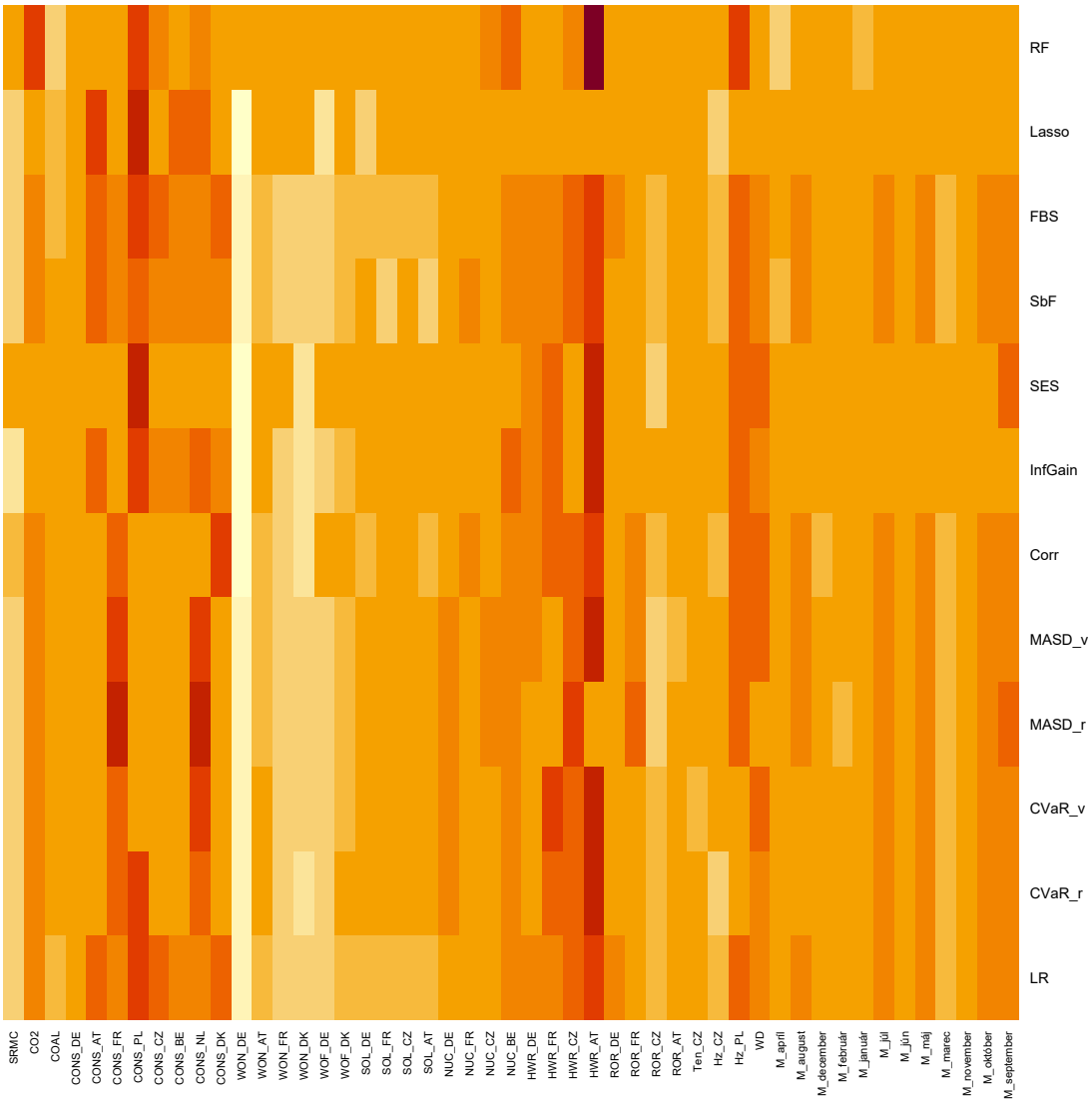
Tabuľka 10: Výsledky predikcie logistickej regresie pre rôzny výber vstupných premenných na dátach z oblasti energetiky.

by bol vhodným kompromisom medzi celkovou úspešnosťou klasifikácie a jednoduchosťou modelu, vzhľadom na interpretáciu, ako aj výpočet predikcie.

Keďže nás okrem výsledkov zaujíma aj aké premenné jednotlivé metódy vyberali a aké β im v logistickej regresii priradili, uvádzame Obrázok heatmapy 20, ktorá nám tieto údaje farebne vizualizuje. Obrázok treba čítať po riadkoch, kde intenzita farby hovorí o tom, akú veľkú váhu β premennej priradila logistická regresia. Čím je farba na obrázku tmavšia, tým sú koeficienty premenných logistickej regresie väčšie oproti ostatným premenným a naopak čím je farba bledšia, tým sú koeficienty menšie. V prípadoch, kedy je farba veľmi svetlá, to znamená záporné váhy odhadovanej β. Najčastejšie sa vyskytujúca farba značí nulové hodnoty β, respektíve tie, ktoré sú blízko nuly. Treba pripomenúť, že nulové hodnoty nie sú spôsobené výpočtom algoritmu logistickej regresie, ale predošlým procesom výberu premenných.

Na základe intenzít koeficientov β, nevieme ani o jednej z premenných povedať, že by bola kľúčovou pre všetky metódy. Vidieť však premenné, ktoré sú dôležité pre väčšinu z metód (tmavá, respektíve veľmi bledá farba). Menovite sú to premenné *CONS_FR,*

CONS_PL, CONS_NL, WON_DE a *HWR_AT*. Môžeme teda predpokladať, že tieto premenné sa najväčšou mierou zasluhujú na výsledkoch predikcií a preto im mnohé z metód na výber premenných priradili vysoké hodnoty β.



Obr. 20: Váhy premenných naprieč výsledkami LR na dátach z oblasti energetiky.

Záver

Cieľom diplomovej práce bolo vysvetliť a implementovať klasifikačný model navrhnutý v článku [9] a následne ho porovnať s bežne využívanými metódami. Ide o klasifikačný model založený na metóde strojového učenia zvanej SVM, ktorý sa dá vnímať ako model na minimalizáciu štrukturálneho rizika, pričom toto riziko je interpretované metódami CVaR, respektíve MASD, čo sú metódy na meranie finančného rizika vo financiách.

V prvej kapitole 1 sme čitateľovi priblížili metódu SVM, vysvetlili jej základné princípy a uviedli problémy, na ktoré sa využíva. Uviedli sme príklad pre klasifikáciu separovateľných (3) aj neseparovateľných dát (6), ako aj jej rôzne obmeny, aby bola použiteľná pre takmer všetky typy klasifikačných úloh.

V druhej kapitole 2 sme uviedli, čo je to finančné riziko a dôvod prečo a ako sa meria. Predstavili sme doteraz najpoužívanejšiu metódu VaR na meranie finančného rizika. V druhej časti kapitoly sme ukázali novšie spôsoby merania finančného rizika formou CVaR a MASD, tak ako ich uviedli v [39] a [9], ktoré zohľadňujú nedostatky modelu VaR. Taktiež je v tejto kapitole definovaná koherentná miera rizika, ktorá je z teoretického hľadiska kľúčová pre predpoklady nášho modelu.

V nasledovnej kapitole 3 uvádzame odvodenie klasifikačných modelov navrhnutých v [9]. Oba pozostávajú z riešenia úlohy lineárneho programovania pri nezáporných váhach a za použitia l-1 normy, ktorá zaručuje nulovosť mnohých váh. To sa neskôr ukazuje, že môže poslúžiť na výber premenných do iných modelov.

V kapitole 4 sme na dátach o diabete otestovali navrhnuté modely CVaR a MASD a porovnali ich výsledky s bežne používanými klasifikačnými metódami (Tabuľka 5). Zistili sme, že obe metódy dávajú porovnateľné výsledky voči zvyšným metódam, pričom v jednej zo sledovaných metrík metóda MASD dosiahla ďaleko najlepšie výsledky.

V úvode kapitoly 5 stručne popisujeme metódy využívané v dátovom pred-processingu, ktoré slúžia na výber premenných. Následne sme porovnávali podobnosť ich výberov na dátach o finančnom audite spolu s metódami CVaR a MASD z článku [9], ktoré tiež majú vlastnosť výberu premenných. V závere sme úspešnosť metód porovnávali pomocou logistickej regresie a to tak, že sme metódu LR nechali zbehnúť pre každý výber premenných určený metódami. Z porovnaní sa ukázalo (Tabuľka 7), že metódy CVaR a MASD vyberali také premenné, ktoré dosahovali výborné predikčné hodnoty v LR.

Poslednú kapitolu 6 sme venovali aplikácii metód CVaR a MASD na dáta z oblasti energetiky v súvislosti s predikciou výroby elektrickej energie plynovými elektrárňami. Dáta bolo na úvod potrebné dôkladne analyzovať a následne sa na nich natrénovali klasifikačné metódy. Z porovnania výsledkov predikcií vyplynulo, že metóda CVaR opäť dosiahla porovnateľné výsledky s ostatnými metódami, avšak metóda MASD v tomto príklade v úspešnosti hlavnej metriky značne zaostávala. Na záver kapitoly, sme sa opäť zamerali na vlastnosť metód CVaR a MASD v súvislosti s výberom premenných, ktoré boli následne použité v logistickej regresii. Výsledky ukazujú, že ani jedna z metód na výber premenných nevybrala takú podskupinu premenných, ktorá by dosahovala lepší výkon v logistickej regresii ako za použitia kompletnej sady premenných. Zároveň však tvrdíme, že použitie metódy CVaR spolu s logistickou regresiou môže byť dobrý kompromis, čo sa týka úspešnosti výsledkov predikcií a počtu vstupných premenných do modelu.

Z meraní v kapitolách 4, 5 a 6 môžeme potvrdiť, že trieda klasifikačných metód založených na finančných mierach rizika CVaR a MASD navrhnutá v článku [9], sa vo viacerých príkladoch ukázala ako vhodnou alternatívou k bežne využívaným klasifikačným metódam, pričom vo viacerých prípadoch dosahovala v jednotlivých metrikách lepšie výsledky ako tieto metódy. Úspešnosť metód CVaR a MASD sa ukázala aj v druhej disciplíne práce s dátami, ktorou bol výber premenných. Tá slúži na zjednodušenie modelov ako aj na skrátenie času výpočtov.

Trieda metód CVaR a MASD už teraz javí možné využitie vo viacerých oblastiach dátovej analytiky. Potenciálne rozšírenie tejto triedy metód siaha do oblastí viacnásobnej klasifikácie, regresie ako aj detekcie outlierov, tak ako sa uvádza v závere článku [9]. Čo sa týka našej práce, možné rozšírenie badať práve v týchto oblastiach, ako aj v optimalizovaní času výpočtu klasifikačných algoritmov CVaR a MASD naprogramovaných v tejto práci (B.1 a B.2).

Zoznam použitej literatúry

[1] Artzner, P. et al.: *Coherent measures of risk*, Mathematical Finance, Vol. 9, No. 3, 1999

[2] Berkelaar, M., et al.: *lpSolve: Interface to 'Lp_solve' v. 5.5 to Solve Linear/Integer Programs*, R package version 5.6.15, 2020, dostupné na internete (04.05.2020): https://CRAN.R-project.org/package=lpSolve

[3] Borboudakis, G., Tsamardinos, I.: *Forward-Backward Selection with Early Dropping*, Journal of Machine Learning Research vol.20 (2019), 1-39, dostupné na internete (06.05.2020): http://jmlr.org/papers/v20/17-334.html

[4] Boser, B. E., Guyon, I. M., Vapnik V. N.: *A Training Algorithm for Optimal Margin Classifiers*, COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory. New York (1992), NY, USA: ACM Press, 144–152

[5] Boyd, S., Vandenberghe, L.: *Convex Optimization*, Cambridge University Press, 2009

[6] Daníelsson, J. et al.: *Subadditivity Re–Examined: the Case for Value–at–Risk*, Working paper, Cornell University, Ithaca, New York, 2005, dostupné na internete (04.05.2020): https://people.orie.cornell.edu/gennady/techreports/VaRsubadd.pdf

[7] European Network of Transmission System Operators for Electricity, dostupné na internete (11.05.2020): https://transparency.entsoe.eu/dashboard/show

[8] Friedman, J., Hastie, T., Tibshirani, R.: *Regularization Paths for Generalized Linear Models via Coordinate Descent*, Journal of Statistical Software 33 (2010), 1-22, dostupné na internete (05.05.2020): http://www.jstatsoft.org/v33/i01/

[9] Gotoh, J., Takeda, A., Yamamoto, R.: *Interaction between financial risk measures and machine learning methods*, Springer, Berlin, 2013

[10] Gotoh, J., Takeda, A.: *A Linear Classification Model Based on Conditional Geometric Score* , Pacific Journal of Optimization vol.1 (2005)

[11] Gareth, J. et al.: *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013

[12] Hall, M. A.: *Correlation-based Feature Selection for Machine Learning*, PhD thesis, The University of Waikato, Hamilton, NewZealand, 1999, dostupné na internete (06.05.2020): https://www.lri.fr/~pierres/donn%E9es/save/these/articles/lpr-queue/hall99correlationbased.pdf

[13] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*, Springer, Stanford, 2008

[14] Haugh, M.: *Quantitative Risk Management*, slajdy z prednášok, Columbia University, New York, dostupné na internete (04.05.2020): http://www.columbia.edu/~mh2078/QRM/RiskMeasures_MasterSlides.pdf

[15] Hod, M. et al.: *Textbook of Diabetes and Pregnancy*, CRC Press, Florida, 2016

[16] Holton, G.A.: *History of Value-at-Risk*, Working Paper, Boston, 2002, dostupné na internete (09.01.2020): http://stat.wharton.upenn.edu/~steele/Courses/434/434Context/RiskManagement/VaRHistlory.pdf

[17] Hooda, N., Bawa, S., Rana, P. S.: *Fraudulent Firm Classification: A Case Study of an External Audit*, Applied Artificial Intelligence, vol. 32, 2018, 48-64, dostupné na internete (06.05.2020): https://doi.org/10.1080/08839514.2018.1451032

[18] Hothorn, T., Hornik, K., Zeileis, A.: *Unbiased Recursive Partitioning: A Conditional Inference Framework*, Journal of Computational and Graphical Statistics Vol.15 (2006), 651-674

[19] Hrbáň J.: *Oceňovanie opcií pomocou neurónových sietí*, bakalárska práca, FMFI UK, Bratislava, 2018, dostupné na internete (04.05.2020): https://opac.crzp.sk/?fn=detailBiblioForm&sid=BCB1C6D618A1774C8CDDC05E7FBF

[20] Charnes, A., Cooper, W. W.: *Programming with linear fractional functionals*, Naval Research Logistics Quarterly, Vol.9, Issue 3-4, 1962

[21] Christopher M. Bishop: *Pattern Recognition and Machine Learning*, Springer, New York, 2006

[22] Intercontinental Exchange, dostupné na internete (11.05.2020): https://www.theice.com/index

[23] Joint Allocation Office, dostupné na internete (11.05.2020): https://www.jao.eu/main

[24] Jović, A., Brkić, K., Bogunović, N.: *A review of feature selection methods with applications*, 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, 2015, pp. 1200-1205, dostupné na internete (06.05.2020): https://ieeexplore.ieee.org/abstract/document/7160458

[25] Kaggle: Your Home for Data Science, dostupné na internete (05.05.2020): https://www.kaggle.com/uciml/pima-indians-diabetes-database

[26] Kaggle: Your Home for Data Science, dostupné na internete (05.05.2020): https://www.kaggle.com/sid321axn/audit-data

[27] Kaviani P., Mrs. Dhotre S.: *Short Survey on Naive Bayes Algorithm*, International Journal of Advance Engineering and Research Development, Vol. 4, Issue 11, 2017

[28] Kuhn, M.: *caret: Classification and Regression Training*, R package version 6.0-86, 2020, dostupné na internete (05.05.2020): https://CRAN.R-project.org/package=caret

[29] Kuhn, M.: *Variable Selection Using The caret Package*, Working paper, 2010, dostupné na internete (06.05.2020): http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.168.1655&rep=rep1&type=pdf

[30] Lagani, V., et al.: *Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets*, Journal of Statistical Software, vol.80 (2017), dostupné na internete (06.05.2020): https://doi.org/10.18637/jss.v080.i07

[31] LeDell, E. et al.: *h2o: R Interface for the 'H2O' Scalable Machine Learning Platform*, R package version 3.30.0.1, 2020, dostupné na internete (05.05.2020): https://CRAN.R-project.org/package=h2o

[32] Li, J. et al.: *Feature Selection: A Data Perspective*, 2016, dostupné na internete (06.05.2020): https://arxiv.org/abs/1601.07996v5

[33] Liaw, A., Wiener, M.: *Classification and Regression by randomForest*, R News 2 (2002), 18-22.

[34] Lu, Z., Zhang, Y.: *Penalty Decomposition Methods for l0-Norm Minimization*, Working paper, Simon Fraser University, Burnaby, Canada, 2010, dostupné na internete (13.05.2020): https://arxiv.org/abs/1008.5372v1

[35] Metz, CH. E.: *Basic principles of ROC analysis*, Seminars in Nuclear Medicine Vol. 8 (1978), 283-298

[36] Meyer, D. et al.: *e1071: Misc Functions of the Department of Statistics*, Probability Theory Group (2019), TU Wien, dostupné na internete (05.05.2020): https://CRAN.R-project.org/package=e1071

[37] R Core Team,: *A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria, 2019, dostupné na internete (13.05.2020): https://www.R-project.org/

[38] Robnik-Sikonja, M., Savicky, P.: *CORElearn: Classification, Regression and Feature Evaluation*, R package version 1.54.2, 2020, dostupné na internete (05.05.2020): https://CRAN.R-project.org/package=CORElearn

[39] Rockafellar, R. T., Uryasev, S. P.: *Optimization of conditional value-at-risk*, Journal of Risk (2000)

[40] S&P Global Platts: S&P Global Platts is a leading provider of energy and commodities information, dostupné na internete (12.05.2020): https://www.spglobal.com/platts/plattscontent/_assets/_files/en/our-methodology/methodology-specifications/european_power_methodology.pdf

[41] Shang, Ch. et al.: *Feature selection via maximizing global information gain for text classification*, Knowledge Based Systems Vol. 54 (2013), 298-309

[42] Shapiro, S. S., Wilk, M. B.: *An Analysis of Variance Test for Normality (Complete Samples)*, Biometrika, Vol. 52 (1965), 591-611, dostupné na internete (05.05.2020): https://www.jstor.org/stable/2333709?seq=1

[43] Schölkopf, B. a kol.: *New Support Vector Algorithms*, Neural Computation 12 (2000), 1207-1245

[44] Student: *The Probable Error of a Mean*, Biometrika Vol. 6 (1908), 1-25, dostupné na internete (05.05.2020): https://www.jstor.org/stable/2331554?seq=1

[45] The European Energy Exchange, dostupné na internete (11.05.2020): https://www.eex.com/en#/en

[46] Toma, V.: *Základy lineárneho programovania*, študijný text, Knižničné a edičné centrum FMFI UK, FMFI UK, Bratislava, 2008, dostupné na internete (05.05.2020): https://zona.fmph.uniba.sk/fileadmin/fmfi/sluzby/elektronicke_studijne_materialy/ip_uk/Zaklady_linerne_programovanie.pdf

[47] Uryasev, S.: *Conditional Value-at-Risk, Methodology and Applications: Overview*, prezentácia, dostupné na internete (13.05.2020): http://www.pacca.info/public/files/docs/public/finance/Active%20Risk%20Management/Uryasev%20-%20Cvar%20Metodology%20And%20Application.pdf

[48] Vanderbei, R. J.: *Linear Programming: Foundaitions and Extensions*, Springer, New York, 2014

[49] Venables, W. N., Ripley, B. D.: *Modern Applied Statistics with S. Fourth Edition.* Springer, New York, 2002, ISBN 0-387-95457-0

[50] Wei, T., Simko, V.:, *R package "corrplot": Visualization of a Correlation Matrix (Version 0.84)*, 2017, dostupné na internete (06.05.2020): https://github.com/taiyun/corrplot

[51] Weihs, C., et al.: *klaR Analyzing German Business Cycles*, Data Analysis and Decision Support (2005), 335-343, Springer-Verlag, Berlin

[52] Wilcoxon, F.: *Individual Comparisons by Ranking Methods*, Biometrics Bulletin Vol. 1 (1945), 80-83, dostupné na internete (05.05.2020): https://www.jstor.org/stable/3001968

Príloha A

A.1 Regularizácia

Ďalšou z možností, vďaka ktorej vieme dospieť k zlepšeniu výsledkov predikcií modelu, je takzvaná regularizácia. Videli sme, že v kapitole 1 príklad (7) sa štrukturálne riziko skladalo z dvoch častí, pričom prvá z nich bol regularizačný člen.

Regularizácia je často používanou metódou na riešenie bi - kriteriálnych optimalizačných problémov. Uvedieme si nasledovný ilustračný príklad (vychádzajúc z [5]):

$$\min_x ||Ax - b|| + \gamma||x||, \tag{A.1}$$

kde parameter $\gamma > 0$ predstavuje penalizáciu za veľkosť normy $||x||$. Takto navrhnutý problém teda optimalizuje dva členy naraz. Výhodou takéhoto riešenia môže byť, že jemné zašumenie v matici A až tak radikálne neovplyvní výraz Ax, čo by sa naopak pri veľkých hodnotách jednotlivých zložiek vektora x mohlo stať.

Zmyslom regularizácie v prípade strojového učenia je predídenie pretrénovania modelu. Istú paralelu medzi príkladom (A.1) a strojovým učením môžeme vidieť v tom, že po natrénovaní modelu na trénovacej sade dát robíme predikciu na testovacej sade. Tá s istými variáciami býva podobná tej trénovacej a preto nízke hodnoty v x nám zabezpečia vierohodné výsledky predikcie. Okrem tejto vlastnosti môže regularizácia slúžiť aj na zjednodušenie modelu. V prípade aplikácie l-1 normy na regularizačný člen nám optimalizačný problém dáva riedke riešenie, čo znamená, že mnohé z optimalizovaných premenných vektora x budú nulové.

A.2 Aproximácia normy

V tejto sekcii si priblížime, akú funkciu má v modeli (14) l-1 norma a prečo práve táto norma dáva takzvané "riedke"riešenie, t.j. mnohé z premenných premietne do nuly. Čerpať budeme z článku [5].

Na úvod si uvedieme všeobecný príklad aproximácie normy bez ohraničení, ako aj konkrétne prípady pre l-1 a l-2 normy ako sú uvedené v [5]. Majme maticu dát $A \in R^{m,n}$, vektor jej odozvy $b \in R^m$ a vektor neznámych $x \in R^n$. Potom riešenie úlohy

$$\underset{x}{\text{minimize}} \quad ||Ax - b||, \tag{A.2}$$

bude približné riešenie systému $Ax \approx b$ za použitia normy $||.||$. Vektor $r = Ax - b$ sa nazýva vektorom *rezíduí* daného problému.

V prípade, že vektor b patrí do stĺpcového priestoru matice A, t.j. $b \in S(A)$, existuje také x, že hodnota účelovej funkcie je nulová a platí $Ax = b$. To isté platí aj v prípade, ak sú rozmery matice A rovné, teda $m = n$, a matica má plnú hodnosť. Aj vzhľadom na naše dáta nás preto bude zaujímať netriviálne riešenie a teda prípad, kedy $m > n$ a zároveň $b \notin S(A)$.

Geometrická interpretácia takéhoto problému je nasledovná: nech $\mathcal{A} = S(A) \subseteq R^m$ a $b \in R^m$, potom

$$\begin{aligned} \underset{u}{\text{minimize}} \quad & ||u - b|| \\ s.t. \quad & u \in \mathcal{A}, \end{aligned}$$

bude znázorňovať projekciu vektoru b do stĺpcového priestoru matice A, pričom bod $u = Ax$, predstavujúci lineárnu kombináciu stĺpcov matice A zabezpečenej vektorom x, je bodom projekcie vektora b.

A.2.1 Aproximácia euklidovskej normy

Pri špecifikácii normy na euklidovskú l-2 normu bude úloha vyzerať nasledovne:

$$\text{minimize} \quad ||Ax - b||_2^2 = r_1^2 + r_2^2 + ... + r_m^2.$$

Daný problém sa dá riešiť aj analyticky, kde optimálne x bude $\hat{x} = (A^T A)^{-1} A^T b$.

A.2.2 Aproximácia jednotkovej normy

Pri špecifikácii normy na jednotkovú (resp. poštársku) l-1 normu bude úloha vyzerať nasledovne:

$$\text{minimize} \quad ||Ax - b||_1 = |r_1| + |r_2| + ... + |r_m|.$$

Takto definovaná norma počíta súčet absolútnych rezíduí. Úlohu je možné formulovať aj ako problém lineárneho programovania

$$\begin{aligned} \text{minimize} \quad & 1^T t \\ s.t. \quad & -t \leq Ax - b \leq t, \end{aligned}$$

kde $t \in R^m$.

A.2.3 Penalizačná funkcia

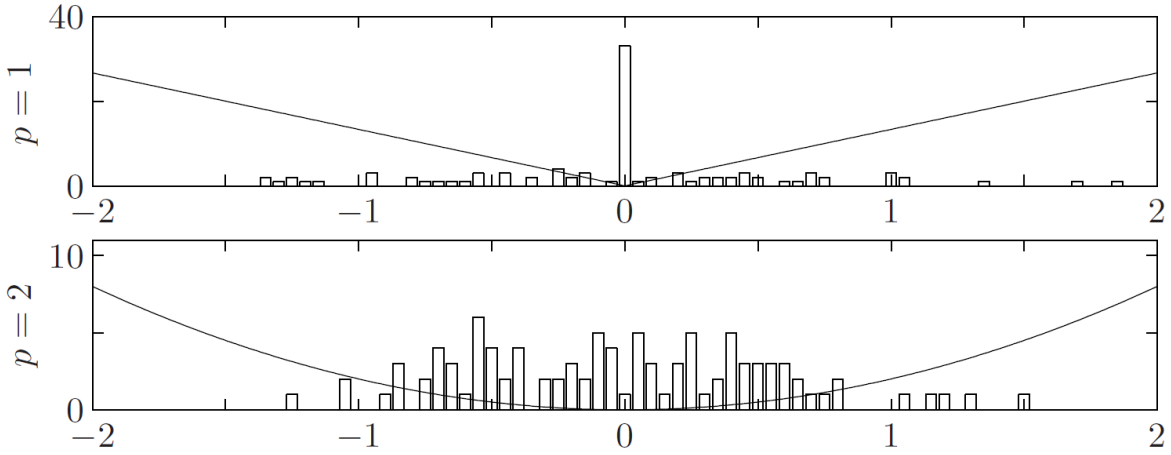
V úlohách aproximácie l-p normy, vieme účelovú funkciu zapísať do tvaru

$$|r_1|^p + |r_2|^p + ... + |r_3|^p,$$

z čoho dostávame separovateľnú a symetrickú funkciu od rezíduí. V takomto prípade účelová funkcia závisí čisto od distribúcie amplitúd rezíduí. Penalizačná funkcia má potom nasledovný tvar:

$$\begin{aligned} \text{minimize} \quad & \Phi(r_1) + \Phi(r_2) + ... + \Phi(r_m) \\ s.t. \quad & r = Ax - b, \end{aligned} \tag{A.3}$$

kde Φ: $R \longrightarrow R$ sa nazýva penalizačná funkcia.



Obr. A.21: Obrázok znázorňujúci využitie l-1 a l-2 normy. Prebraté z [5].

Ako sme už písali na úvod tejto podkapitoly, chceli sme vysvetliť dôvod výberu l-1 normy autormi z článku [9] do modelu z kapitoly 3 v príklade (14). Ako môžeme vidieť z

Obrázku A.21, výber normy má rozhodujúci vplyv na výslednú hodnotu váh jednotlivých rezíduí. V prípade kvadratickej funkcie vidíme, že pre nízke hodnoty rezíduí má aj funkcia veľmi nízke hodnoty, preto nevadí, že im úloha priradí nejakú váhu, lebo na výslednej hodnote penalizačnej funkcie sa podieľajú len malou mierou. Keď si však vezmeme funkciu absolútnych hodnôt vidíme, že aj malá chyba pridáva do penalizačnej funkcie jej absolútnu hodnotu, preto má takáto funkcia tendenciu tlačiť mnoho hodnôt do nuly. Inak povedané, funkcia $\Phi(r)$ je akousi mierou našej averzie voči vysokým, resp. nízkym hodnotám rezíduí pri vstupnej premennej r.

Príloha B

B.1 Kód metódy CVaR v softvéri R

```
CVaR <- function(data, beta, prob){
  # data - vstupne data urcene na natrenovanie modelu
  # beta - hodnota parametra vstupujuceho do modelu
  # prob - vektor pravdepodobnosti, vstupujuci do modelu

  y <- data$y
  X <- data%>%select(-y)%>%as.matrix()
  p <- ncol(X)
  m <- nrow(X)

  M <- y*X
  e1 <- rep(1,m)
  e2 <- rep(0,m+p+4)
  e2[(m+3):(m+p+2)] <- 1
  B <- cbind(e1, -e1, diag(m), M, -y, y)
  B <- rbind(e2, B)

  ### Premenne do LP modelu
  direction <- 'min'
  objective.in <- c(1, -1, 1/(1-beta)*prob, rep(0,p+2))
  const.mat <- B
  const.dir <- c('=', rep('>=', m))
  const.rhs <- c(1, rep(0, m))

  ### Uloha LP
  linp <- lp(direction, objective.in, const.mat, const.dir, const.rhs)

  ### Vystupne premenne
```

```
  w <- linp$solution[(m+3):(m+p+2)]
  b <- linp$solution[(dim(B)[2]-1)] - linp$solution[(dim(B)[2])]

  return(list(w=w, b=b, Beta = beta))
}
```

B.2 Kód metódy MASD v softvéri R

```
MASD <- function(data,lambda,prob){
  # data   - vstupne data urcene na natrenovanie modelu
  # lambda - hodnota parametra vstupujuceho do modelu
  # prob   - vektor pravdepodobnosti, vstupujuci do modelu

  y <- data$y
  X <- data%>%select(-y)%>%as.matrix()
  p <- ncol(X)
  m <- nrow(X)

  temp1 <- y*X
  temp2 <- sum(prob*y)
  e1 <- rep(1,m)
  Mw <- temp1 - outer(e1, prob)%*%temp1
  Mb <- -y + temp2
  e2 <- rep(0,m+p+2)
  e2[1:p] <- 1
  B <- cbind(Mw, Mb, -Mb, diag(m))
  B <- rbind(e2, B)

  ### Premenne do LP modelu
  direction <- 'min'
  objective.in <- c(-t(prob) %*% temp1, temp2, -temp2, lambda*prob)
```

```
  const.mat <- B
  const.dir <- c('=', rep('>=', m))
  const.rhs <- c(1, rep(0, m))

  ### Uloha LP
  linp <- lp(direction, objective.in, const.mat, const.dir, const.rhs)

  ### Vystupne premenne
  w <- linp$solution[1:p]
  b <- linp$solution[(dim(B)[2]-1)] - linp$solution[(dim(B)[2])]

  return(list(w=w, b=b, Lambda=lambda))
}
```